A comparison of next-generation sequencing protocols for microbial profiling

Thesis · February 2015

DOI: 10.13140/RG.2.2.25947.16165

1 author:

Richard Fong
Massey University
24 PUBLICATIONS **101** CITATIONS

SEE PROFILE

Institute of Fundamental Sciences

A comparison of next-generation sequencing protocols for microbial profiling

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Genetics

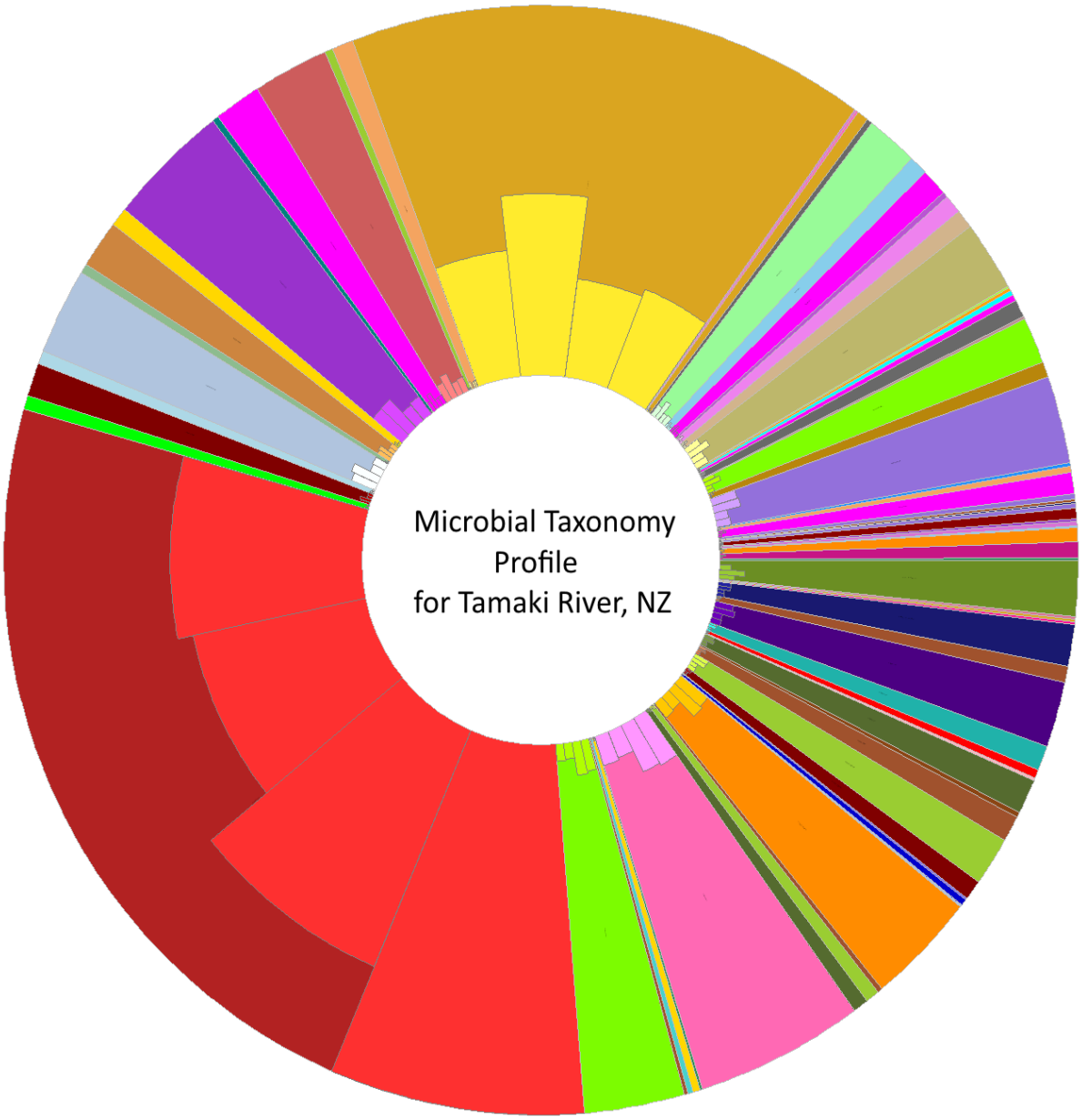
By

Yang Fong (Richard)

2016

Massey University, Palmerston North

New Zealand



One of the responsibilities faced by the Environmental Genome Project is to provide the science base upon which society can make better informed risk management decisions.

-Samuel Wilson-

Abstract

The introduction of massive parallel sequencing has revolutionized analyses of microbial communities. Illumina and other Whole Genome Shotgun Sequencing (WGS) sequencing protocols have promised improved opportunities for investigation of microbial communities. In the present work, we compared and contrasted the findings from different NGS library preparation protocols (Illumina Nextera, Nextera-XT, NEXTFlex PCR-free and Ion-Xpress-400bp) and two sequencing platforms (MiSeq and Ion-Torrent). Short reads were analysed using the rapid database matching software PAUDA and visualization software MEGAN5, which provides a conservative approach for taxonomic identification and functional analyses. In analyses of a Tamaki River water sample, biological inferences were made and compared across platforms and protocols. For even a relatively small number of reads generated on the MiSeq sequencing platform important pathogens were identified in the water sample. Far greater phylogenetic resolution was obtained with WGS sequencing protocols than has been reported in similar studies that have used 16S rDNA Illumina sequencing protocols. TruSeq and Nextera-XT sequencing protocols produced similar results. The latter protocol offered cheaper, and faster results from less DNA starting material. Proteobacteria (alpha, beta and gamma), Actinobacteria and Bacteroidetes were identified as major microbial elements in the Tamaki River sample. Our findings support the emerging view that short read sequence data and enzymatic library prep protocols provide a cost effective tool for evaluating, cataloguing and monitoring microbial species and communities. This is an approach that complements, and provides additional insight to microbial culture “water testing” protocols routinely used for analysing aquatic environments.

Acknowledgement

There are many people I would like to express my gratitude and cordial thanks in helping me out in preparing my Master's Thesis. This dissertation would not have been possible without your support and strong collaboration between different academia backgrounds.

To my lovely wife, I know I could not have done this without your constant encouragement and great patience at all times. I know that during this challenging period, you have been understanding and have given me much help, showered me with your love and support and there are no words to express my appreciation for having you by my side. To my family members; my parents, in-laws and brother, I would like to send my appreciation and would like to say thank you for being patient and for your unequivocal moral support during my master degree course. To my mentor Trish McLenachan, without you my thesis would be incomplete! Thank you for your valuable guidance, advice and patience with my writing. I send you immeasurable and deepest gratitude for your contribution in making this study worthwhile and possible.

To my principal supervisor Professor Peter-James Lockhart, co-supervisor Professor Nigel French and my bioinformatics co-supervisor Dr Patrick Biggs, I would like to thank you for being very supportive, understanding, patient and for providing precious academic and technical advice. To my principal supervisor Peter, I would like to thank you for your guidance throughout my course especially in having faith in me to finish my dissertation on time. Despite project challenges you have continued to encourage me with your knowledge and invaluable thoughts both on an academic and personal level. Special thanks to Dr Patrick Biggs for your continual support and enlightenment regarding bioinformatics analyses.

For financial support, I would like to acknowledge and thank the Institute of Veterinary, Animal and Biomedical Sciences (IVABS), Institute of Fundamental Sciences (IFS), Ministry of Health (MOH), Protozoal Research Unit (PRU, Hopkirk Research Institute). Lastly to all my fellow friends and colleagues, you guys kept me going and most importantly were understanding and patient with my workload and study. You guys are awesome!

Table of Contents

List of Acronyms

%	Percent
°C	Degrees Celsius
µl	Microlitre(s)
µM	Micromolar
100 PE	2 x 100 base pair paired-end read
150 PE	2 x 150 base pair paired-end read
250 PE	2 x 250 base pair paired-end read
300 PE	2 x 300 base pair paired-end read
A	Adenine
A260	Nanodrop absorbance at 260 nanometres
A280	Nanodrop absorbance at 280 nanometres
AFLP	Amplified Fragment Length Polymorphism
ATL	A-Tailing Mix
ATM	Amplicon Tagment Mix
ATP	Adenosine Triphosphate
BAM	Binary Alignment Matrix
BGI	Beijing Genomics Limited
BIPES	Illumina Multiplexed Paired-end Sequencing Adapter
BLAST	Basic Local Alignment Search Tool
bp	Base pair(s)
C	Cytosine
CCD	Charge-coupled Device
cDNA	Complementary Deoxyribonucleic Acid
contig	Continuous Sequence
CTA	A-Tailing Control

CTE	End-Repair Control
CTL	Ligation Control
ddNTP	Dideoxy Nucleotide Triphosphate
dH_2O	Distilled Water
DNA	Deoxyribonucleic Acid
dNTP	Deoxy Nucleotide Triphosphate
ds	Double Stranded
EB	Elution Buffer
eDNA	Environmental Deoxyribonucleic Acid
EDTA	Ethylenediamine Tetra-Acetic Acid
emPCR	Emulsion Polymerase Chain Reaction
ERP	End-Repair mix
EtBr	Ethidium Bromide
E-value	A parameter that describes the number of expected matches when searching a sequence database of a particular size and composition
FC	Flowcell
fq	Fastq File Format
g	Gram(s)
G	Guanine
Gb	Gigabytes
gDNA	Genomic DNA
HiFi	High fidelity enyzme
HMW	High Molecular Weight
HT1	Hybridization Buffer
Inc.	Incorporated
Indel	Small Insertion or deletion
ISFET	Ion Sensitive Field Effect Transistor

KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	Lowest Common Ancestor
LIG	Ligation Mix
Log^{10}	Logarithm to the base 10
M	Molar
Mb	Megabytes
MDA	Multiple Displacement Amplification
MEGAN	Metagenome Analyzer
MGS	Massey Genome Service
min	Minute(s)
ml	Millilitre(s)
mm	Millimetre(s)
mM	Millimolar
MPSS	Massive Parallel Signature Sequencing
mRNA	Messenger Ribonucleic Acid
mtDNA	Mitochondrial Deoxyribonucleic Acid
ng	Nanogram(s)
NGS	Next-generation Sequencing
No	Number
NPM	Nextera PCR Master Mix
NPS	Non-point Source
nt	Nucleotide
NT	Neutralize Tagment Buffer
NZGL	New Zealand Genomics Limited
OTU	Operational Taxonomic Unit
PAUDA	Protein Alignment Using a DNA Aligner

PCoA	Principal Coordinate Analysis
PCR	Polymerase chain reaction
pDNA	Pseudo DNA
PE	Paired-end
PGM	Personal Genome Machine
PhiX	Bacteriophage PhiX174
PMM	PCR Master Mix
pmol	Picomole(s)
PPC	PCR Primer Cocktail
PP_i	Pyrophosphate
Q_{10}	Phred Quality Score 1 error in 10
Q_{20}	Phred Quality Score 1 error in 100
Q_{30}	Phred Quality Score 1 error in 1000
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
Q-score	Phred Quality Score
RNA	Ribonucleic Acid
rpm	Revolutions per Minute
rRNA	Ribosomal Ribonucleic Acid
RSB	Resuspension Buffer
RTA	Real-Time Analysis
s	Second(s)
SAM	Sequence Alignment Map
SBS	Sequencing by Synthesis
SCIMM	Sequence Clustering with Interpolated Markov Models
SEED	Database infrastructure for comparative genomics in MEGAN5 software

SMRT	Single Molecule Real Time
spp.	Species
SPRI	Solid Phase Reversible Immobilization
ss	Single Stranded
STL	Stop Ligation Buffer
T	Thymine
TAE	Tris-Acetate EDTA buffer
TAP	Taxonomic Assignment Pipeline
Taq	*Thermus aquaticus*
TB	Tuberculosis
TD	Tagmentation Buffer
TE	Tris EDTA Buffer
V	Volts
WGS	Whole Genome Shotgun Sequencing

List of Figures

List of Tables

1 Introduction

1.1 Background

The biosphere consists of three important elements: earth, air and water (Lin et al., 2003; Whitman et al., 1998) where each element provides rich habitats for living organisms which interact with each other in complex and diverse ways (Press, 2007). Yet, how much do we really understand about these organisms and their interactions? Microorganisms or microbes are classified as living entities smaller than about 100 µm in size (Kirchman, 2012). They consist of prokaryotes (unicellular organisms comprising eubacteria and archaea), eukaryotes (multicellular organisms including fungi and protists) and viruses (Kirchman, 2012). The study of microbial ecology is challenging due to the extreme variation of the diversity of microbial life and their habitats on earth. A further complicating factor is their high levels of abundance. In one recent study of marine microorganisms for example, one litre of collected water was shown to contain thousands of microbes that thrive, co-exist and interact with other species in a community (Azam et al., 2007). Insight into the composition and dynamics of aquatic populations can be gained using different approaches: microscopy, microbial culture and biochemical techniques, including DNA sequencing and phylogenetic analyses.

Microbes are important organisms because they produce many foods (i.e. cheese, wine, yoghurt) and act at the foundation of the 'food web (food-chain supplier)' within ecosystems. The world is covered with 70% of water where only 3% is classified as fresh water and about 0.5% of this is drinkable. The remaining 2.5% is stored frozen in glaciers of the North (Arctic) and South (Antarctica) Poles (Loucks, 2005). Thus, it is very important to understand what is needed for the conservation of such important ecosystems. In freshwater ecosystems (e.g. lakes, rivers and aquifers) many of the food chains and webs require an abundant mixture of aquatic and soil microbes to sustain and support life. The interactions between microorganisms in aquatic ecosystems contribute to the 'biogeochemical' cycle (Newton et al., 2011). Here, the word biogeochemical refers to the nutrient cycles (carbon, methane, nitrogen, oxygen, phosphorus and sulphur) facilitated by microbes in freshwater ecosystem (Falkowski et al., 2008). For example, in freshwater ecosystems (the aquatic biosphere), microbes (i.e. phytoplankton) interact with numerous macroscopic plants to convert carbon dioxide to organic materials and also to produce inorganic nutrients to sustain growth and biomass production (Falkowski et al., 2008; Kirchman, 2012). Microorganisms were first

observed in the seventeenth century and have been studied extensively ever since with invention of the compound microscope in 1665. There were several historical notes on the study of microorganisms prior to this, but in 1665, the invention of the single lens microscope with 200-fold magnification and illumination system by Robert Hooke accelerated the research of microorganisms. About the same time, another microbiologist Anton van Leeuwenhoek discovered a group of organisms known as the bacteria. He also viewed '*spermatozoa*' for the first time and documented the first discovery of live bacteria known as '*animalcules*' (tiny animals) as well as collections of '*infusoria*' (aquatic creatures) in freshwater ponds (Van Zuylen, 1981). This was followed up by further experimentation by Edward Jenner and Robert Koch on pathogenic strains of microorganisms such as *Bacillus anthracis* (anthrax), *Mycobacterium tuberculosis* (TB) and *variola* virus (small pox) that cause serious diseases in the human population. The concept of microorganisms was further validated by Louis Pasteur in 1854 with the discovery of microorganisms in his sterilized broth after exposure to contaminated air. This led to the fundamental cell theory that 'all living things come only from pre-existing living entities'. The discovery of "pasteurization" greatly aided the decontamination process of many food and beverages during the eighteen century. Over the years both Edward Jenner and Robert Koch successfully founded and revolutionized the field of immunology by developing vaccines for immunization against the above diseases with a high eradication success rate. Most recently, developments and improvements in applied microbiology have given rise to new methodologies for investigating microbial biodiversity and its properties of all elements in the biosphere/ecosystem (Frias-Lopez et al., 2008). In particular, it is now possible to study microorganisms based on the study of genetic material recovered directly from environmental samples. As described more formally below, this rapidly growing discipline is known today as 'metagenomics'. The discipline brings two great advances to environmental microbiology; (1) methods do not rely on culturing the microbes and (2) availability of information on the composition and functioning of microbial communities.

1.2 Common communicable diseases in New Zealand

New Zealand is an environmnetally diverse country with a rich agricultural heritage. However, recent expansion and urbanization in rural areas has contributed to environmental pollution that may disrupt water sustainability in the near future. Agricultural waste is also of increasing concern particularly as a result of the growing awareness of Non-Point Source (NPS) pollution, i.e. polution appearing at a distance from its source caused by rainfall moving over and through contaminated farm ground/soil (Zhang et al., 2011). Organic and chemical pollutants such as abiotic and biotic wastes (pesticides, ammonia, fertilizers, animal waste) from industry facilities, farmland, barnyards and feedlots can significantly contaminate and impact nearby rivers, aquifers and lake ecosystems. Animal waste 'runoff', is thought to be one of the major contributors to NPS pollution and often for this situation there are no appropriate management systems to uphold and maintain the rules and regulations developed for the prevention and control of water pollution. One consequence of animal and fertilizer waste runoff into river ecosystems is "hypertrophism" where high levels of nutrient-rich organic matter encourage the growth of many opportunistic waterborne pathogens including *Vibrio cholerae* which can lead to serious gastrointestinal disease (Gotuzzo et al., 1994; Ongley, 1996).

In many low income countries, waterborne diseases such as cholera are often associated with environmental contamination resulting from poor sanitation (Marquez, 2002). These illnesses are also known to be associated with low socioeconomic status where poverty, economic and health inequalities along with the occurrence of natural disasters can contribute to poor sanitation and cholera outbreaks. (Gotuzzo et al., 1994; Telesmanich et al., 2011). In New Zealand, our freshwater resources such as rivers, lakes and many reservoirs are at similar risk due to contamination from pastoral farming, dairy conversions and natural disasters i.e. earthquakes. One of the most significant pollutants in NZ is nitrate contamination and according to a recent survey by the Ministry of the Environment in 2008, more than one third (39%) of New Zealand groundwater sites have levels of nitrate above the recommended level and this is increasing at an alarming rate, due to leaching of fertiliser and stock effluent from farmland (Daughney et al., 2009). This nutrient-rich waste product provides an ideal environment for many aerobic and anaerobic pathogens to flourish and these can significantly impact on agriculture livestock and public health. In view of this, each year the Ministry of Health in New Zealand (MoH) spends a considerable amount of money monitoring public

health risks. For example, a recent publication by Baker and colleagues in 2013, on systemic disease of close contact infectious diseases (CCID) in New Zealand, highlighted the problem of health inequality and disease risk burden in Maori communities due to household crowding. The report confirmed that at least one in 10 hospital admissions for infectious disease in New Zealand including pneumonia, meningococcal, tuberculosis, and measles were due to an overcrowded household (Baker, M. G. et al., 2013). This study highlighted the importance of good living standards in a household such as heating, insulation and positive air-flow ventilation in reducing respiratory illness. Another report by MoH investigator Andrew Ball (2006), estimating the burden of waterborne disease in New Zealand showed the relationship of drinking water-quality and waterborne gastro-intestinal disease (GID) in New Zealand. According to that health report there are about 17,000 total notified cases of water-borne gastroenteritis reported in New Zealand every year and majority of them are caused by several opportunistic communicable disease agents such as *Campylobacter*, *Salmonella*, *Shigella*, *Yersinia* and toxigenic *E.coli*, protozoa (*Cryptosporidium* spp. and *Giardia* spp) and viruses (*enterovirus*). Furthermore, during the period between 2001 to 2005, New Zealand had a total of 724 confirmed waterborne-gastroenteritis outbreaks and 84 of these reported cases required hospitalization. It is estimated that there has been an average of 145 outbreaks per year from 2001 to 2005; and compared to other countries, New Zealand is considered a high-risk country for water-borne related disease (Andrew, 2006).In his report, Andrew concluded the importance of having a water-treatment plan and that this should be carried out extensively for all drinking water-supplies in New Zealand especially in accordance with WHO guidelines for safe-drinking water. This is less than 1 *E.coli* colony forming unit /100mL of water (Andrew, 2006).

More recently, an investigation lead by Richard Hall and his colleagues from Environmental, Science and Research (ESR) focused on using metagenomic techniques to investigate, identify and catalogue the causative agents responsible for lung cancers in meat-workers. To investigate the air quality in the workplace, two sets of nine aerosol samples were pooled from two different sites; cattle and sheep slaughterhouses respectively (Hall et al., 2013). DNA from aerosol samples were collected from the workers via a special breathing filter apparatus which was then extracted and sequenced using Illumina HiSeq2000 sequencers. The sequencing generated a total of 332,677,436 (cattle slaughter house) and 250,144,492 (sheep slaughter house) sequencing reads of ~85bp (after trimming from 100bp) in length for

bioinformatics analysis. In their bioinformatics analysis, they discovered that sequences from the cattle slaughterhouse had a higher exposure rate and presence of WU polyomavirus and human papillomavirus 120, and that these microorganisms could not have originated by cross-contamination between samples from different sites. Although they found no evidence that exposure to several chemicals used for slaughtering caused lung cancer, they nonetheless discovered that there was a correlation between having a higher count of WU polyomavirus and human papillomavirus 120 with individuals inhaling the bio-aerosol in the slaughterhouse, and who had lung cancer. Although there was no discovery of a causative agent, their findings suggested that there is an occupational risk that requires more attention and further investigation. They also concluded that the metagenomic techniques adopted in the study could be applied for the investigation of microbes in other types of environmental samples such as water and soil ecosystems (Hall et al., 2013).

1.3 Overview of Metagenomics

1.3.1 What is metagenomics?

Metagenomics or environmental genomics is used as a technique for the recovery of genetic material directly from an environmental sample without culturing (Ghazanfar et al., 2010). Genes collected from the environment are analysed to provide information on the genetic, physiological and biochemical interactions of microorganisms living in the sampled environment (Handelsman, 2004). Metagenomic studies began in the early 1980s with pioneering work on rRNA and rDNA genes led by Norman Pace and colleagues at the University of Illinois (Pace et al., 2012). Studies on other genes soon followed. This has included characterization of nuclear encoded 5S rRNA, 18S rRNA and 28S rRNA genes, mitochondrial encoded cytochrome oxidase and mitochondrial 12S genes, as well as chloroplast encoded rbcL, matK and rpl16 genes (Amit Roy, 2014). In fact the word "META" in Greek literally means "beyond" and addition of "genomics" refers to the study of more than one gene (Gilbert et al., 2011). Recently, the word "MEGA" has also been used as an alternative name representing metagenomics sequencing data output from next generation sequencing (NGS) where it typically involves computational analyses (Handelsman, 2005).

Recent improvements in NGS protocols have increased data volume at reduced cost. These developments increase the efficiency of detecting, evaluating, cataloguing and monitoring microbial biodiversity in environmental samples. They provide a powerful means for

undertaking public health assessments such as needed for drinking-water quality management. It is in this context, and with consideration for the importance of surveillance for pathogens, that the present study has been undertaken. This thesis reports a comparative analysis of microbial profiles obtained using different sequencing protocols and NGS (Illumina and Ion-Torrent) platforms.

1.3.2 Types of microbial sequencing methods

Generally there are four types of questions asked in metagenomic investigations "Who is out there?", "How many are there?", "What are they doing?" and "How do they compare?" Four approaches have been generally adopted by researchers. These are (1) amplicon sequencing, most often for 16S rRNA gene regions, (2) low coverage shotgun whole genome sequencing, (3) high coverage whole-genome sequencing involving *de novo* assembly and (4) microbial transcriptome sequencing (RNA-Seq).

16S rRNA/rDNA sequencing

Prior to the analyses of RNA and DNA, both prokaryotes and eukaryotes were classified on their phenotypic characteristics and placed into a taxonomic hierarchy comprising kingdoms, phyla, classes, orders, families, genera and species (Woo et al., 2008). rDNA sequence analyses have been important in the taxonomic revision of microorganisms within this hierarchy (Pace et al., 2012). Furthermore, rDNA sequencing has provided a culture-independent approach for obtaining a more informative representation of microbial diversity. In prokaryotes, rDNA studies have included analyses of three types of conserved rRNA genes: 5S, 16S and 23S rRNA regions. Most published studies have characterised 16S rDNA genes. The potential of the 16S rRNA gene for molecular systematic investigations arises because it is a ubiquitous housekeeping genetic marker that is essential for life with stem regions that are highly conserved across most prokaryotes (eubacteria and archaebacteria). These conserved regions, which can be targeted with "universal primers" flank nine variable loop (V1-V9) regions that vary between taxa. Phylogenetic analyses of the variable gene regions has provided taxonomic resolution at the level of genus, and sometimes also at the level of species (Janda et al., 2007). This has proven useful for both clinical and scientific research (Chakravorty et al., 2007). The highly conserved regions flanking the V regions can be easily targeted by universal PCR primers (Baker, G. C. et al., 2003; Mccabe et al., 1999), and sequencing protocols for ABI3730 and NGS platforms are well established. However,

numerous issues have also arisen indicating the limitations of 16S rDNA sequencing in metagenome studies. These include the realisation that universal primers are not universal for all bacteria and that amplification biases can occur during PCR (reviewed in Wang et al. 2013). In relation to this potential problem, a recent study conducted by Anna et al., (2013) evaluated "in-silico" (i.e. by computer simulation) a total of 175 primers and 512 primer pairs for 16S rDNA based on overall sequencing coverage and taxa in a non-redundant nucleotide rRNA SILVA dataset (SSURef 108NR, (Klindworth et al., 2013). Their analysis showed that only a total of 122 out of 512 primer sets gave confidence scores indicating greater than 50% taxonomic coverage for archaea, bacteria and eukaryotes (Klindworth et al., 2013). They concluded that only 10 general primer sets could be recommended as broad range primers and that the primers chosen should consider first the anticipated microbial diversity. They recommended that such analysis be first evaluated prior to actual amplification to reduce time, cost and bias in microbial diversity study (Klindworth et al., 2013). Of most concern is that relative abundance information of taxa can also be misled by universal primer bias – i.e. templates that better match primers are preferentially amplified, and thus can appear more abundant (Wang, J. et al., 2013). Furthermore even, if multiple variable regions are targeted, phylogenetic resolution might not be obtainable to provide species specific information, which might be required for pathogen identification (Singh et al., 2012). Such findings have encouraged researchers to investigate other molecules and approaches for metagenomics studies.

Low coverage whole genome shotgun sequencing

One early investigation led by Manichanh et al. (2008) analysed approximately 10,010 random ABI3730 sequence reads (RSRs) generated from a cloned DNA library of human faecal samples (Manichanh et al., 2008). They compared this approach with the 16S sequencing analysis method for estimating the biodiversity of their metagenomic library. They demonstrated that using RSR sequence analysis could be a faster and cheaper alternative to 16S amplicon sequence analysis. Both methods were subjected to the same computational pipeline "TAP" (Taxonomic Assignment Pipeline), and searches using "BLASTN" returned a similar and comparable result between both methods (Manichanh et al., 2008). To further verify the efficiency of their method, they downloaded a published Sargasso Sea dataset and subjected it to the same TAP analysis. They found that the diversity pattern was comparable and similar to the 16S approach (Manichanh et al., 2008). This result

highlighted the similarity in biodiversity and consistency between both techniques (RSR and 16S analyses) and the authors concluded both methods were reliable. Their findings suggested that at least for relatively small data sets, RSR analysis provides an alternative protocol to 16S rDNA sequence analysis (Manichanh et al., 2008).

A generalisation of this approach involves whole genome shotgun sequencing, in which DNA from an environmental sample is fragmented and fragments sequenced at random. NGS technology can be applied in this situation to provide a whole-community analysis of the total microbial community structure and diversity in the sample. This approach also provides a means for carrying out functional analysis of proteins present and/or analysis of genes expressed by the microbial community. However, the approach requires a higher level of sequencing coverage than the study conducted by Manichanh et al. (2008) in order to identify the minority community members in a metagenomics sample. The authors also reiterated the importance of having a strategic plan for computational analysis as the large numbers of sequence reads required were a potential bottleneck for many NGS applications (Manichanh et al., 2008).

In low coverage whole genome shotgun sequencing, long fragments of DNA are broken into smaller pieces of DNA (less than 1kb) and these are then processed and sequenced in parallel. The Illumina sequencing platform will produce millions of reads, some of which overlap and can be aligned to produce "contigs" before being matched to databases. Alternatively the short reads can be directly matched to database references. The challenge for shotgun metagenomic sequencing is not the sample preparation and sequencing itself but application of the complex algorithms required to identify and interpret the digital metagenome information. With the rapid expansion of current NGS approaches and sequencing data, computational tools such as the MG-RAST, metaBEETL, PAUDA, LAMBDA, DIAMOND and MEGAN and have been recently developed to help catalogue the metagenome data for taxonomic classification and functional analyses (Ander et al., 2013; Buchfink et al., 2015; Hauswedell et al., 2014; Huson et al., 2007; Huson, D. et al., 2014; Meyer et al., 2008).

A recent publication by (Hasman et al., 2014) on the rapid identification and characterization of microbial diversity from clinical samples using shotgun metagenomic sequencing investigated thirty-five random urine samples from patients suspected of having urinary tract infections. The study included a comparison of a conventional microbiology culture-based

method with the whole genome shotgun sequencing of cultured isolates and clinical urine samples. Thirty five samples were spread, and bacteria cultured, on blood agar plates. These were incubated overnight under aerobic conditions, and from the plates at least one colony was identified in the traditional manner and from this DNA was isolated for whole genome shotgun sequencing. DNA was also extracted directly from the infected urine clinical samples for direct whole metagenomics sequencing. The cultures produced 19 different isolates with eight samples identified biochemically and morphologically as *E. coli*, six samples as *Enterococcus* spp, two samples as *Proteus* spp and one as *Staphylococcus* spp (Hasman et al., 2014). These results were corroborated by whole-genome sequencing from the cultures and direct sequencing of DNA extracted from the urine samples. The authors concluded that the sequencing provided reliable information and drastically reduced the cost and time required for diagnosis.

High coverage whole genome shotgun sequencing

At high levels of sequence coverage *de novo* assembly and annotation of genomes can be conducted to provide higher level functional information. Unlike community profiling, sequencing more complete whole microbial genomes provides more information on the functional characteristics and activities of a microbial community. The recent introduction of Illumina HiSeq X-Ten sequencing system is designed specifically for the high-throughput ultra-deep sequencing needed for population-scale genomics studies. This approach can aid discovery and detection of intraspecific variation, post translational modifications, as well as enable researchers to discover novel genes for applications in the biotechnology industry (Lorenz et al., 2005). It has the potential to provide more information for the diagnosis and surveillance of pathogens that occur in low abundance which are responsible for infectious disease.

The study by Lecuit et al. (2014) highlighted the use of NGS technology for the diagnosis of several major bacterial and viral genomes (*Treponema palidum, Mycobacterium leprae,* Hepatitis A, B, C and E viruses). They concluded such an approach will replace the aging associative diagnostic methods (pathogen specific-PCR) currently being used for the discovery of pathogens (Lecuit et al., 2014). They also emphasized the importance of establishing metagenomic diagnostic tools for the qualitative analysis of sequences from human, animal or environmental samples (Lecuit et al., 2014). In another study, Barzon et al.

(2012) have also supported the application of NGS technology for the detection of emerging viral infections in diagnostic virology. These authors effectively utilized an ultra-deep sequencing method (involving *de novo* assembly) for the diagnosis of hard-to-clone viral pathogens together with their unique drug resistance genes (Barzon et al., 2011). They suggested that the use of NGS technology in the current clinical virology diagnostic setting could broaden the detection of disease-associated viruses and enable the discovery of novel human viruses including cancer-inducing viruses such as HIV and Hepatitis C (Barzon et al., 2013).

Random shotgun sequencing of cDNA

Transcriptome sequencing, whole-transcriptomics sequencing (WTS) or RNA-(cDNA) sequencing (RNA-Seq) is an approach where total RNA retrieved from an environmental sample is converted to cDNA, sequenced and analysed to determine the functional activity of microbial populations. In sequencing cDNA, the RNA-Seq data includes the different populations of total RNA such as mRNA, miRNA, tRNA, siRNA, and sRNA (non-coding RNA) (Wang, Z. et al., 2009). Sequence analyses of environmental RNA can help determine whether different microbes in the same community are active in the same metabolic pathways. It can also show the extent to which microbial metabolisms differ between environments. These issues are increasingly being considered by researchers in analyses of metagenomic data. Such computational metabolic pathway analyses can provide vital clues to our understanding of how microbes interact with each other within ecosystems. cDNA sequencing can provide precise measurement of the level of transcripts, and provide information on virulence factors and other markers of pathogenicity which are important for epidemiological studies.

Combining cDNA and DNA sequencing

For a comprehensive metagenomics survey of microbial communities that maximizes information of taxonomic description and ecological function, both DNA and cDNA can be sequenced. Such a study was undertaken by Yu et al. (2012). These researchers combined both approaches to survey microbial activities and composition in an activated sludge community, in a wastewater treatment plant in Hong Kong. Total DNA and mRNA (RNA-Seq) were sequenced to a sequencing depth of 2.4 Gb (100x coverage) on the Illumina HiSeq2000 platform (Yu et al., 2012). A total of 26,597,304 DNA clean reads (100 bp) and

27,999,804 cDNA (RNA-Seq, 90 bp) were generated from the HiSeq run with a total combined of 53 million clean reads from both DNA and RNA sequencing datasets. Further taxonomic analysis via MG-RAST and the SILVA SSU-reference (16S/18S rDNA) database revealed that both DNA and RNA sequencing datasets had similar annotations. In a sample 'activated' sludge collected from a wastewater treatment plant, the microbial community was dominated by Proteobacteria (22.35%), Actinobacteria (15.03%), Bacteroidetes (5.72%), and Firmicutes (3.22%) (Yu et al., 2012). In addition, functional analyses of the transcriptome 16S/18S rDNA and MG-RAST Genbank database, revealed nitrifying genes were expressed at a relative higher level. Similarly, ammonia monooxygenase and hydroxylamine oxidase enzymes, which contributes to nitrification activity, were also highly expressed. This strong nitrification activity could be assigned to a large population of facultative anaerobic bacteria from the Proteobacteria phyla (*Nitrosomonas* and *Nitrosospira* spp) (Yu et al., 2012). This study highlights, the value of taxonomic analyses with meta-transcriptome data. Such analyses provide a more informed understanding of the microbial community (Yu et al., 2012).

Summary of current metagenomic approaches

The recent advances in NGS technology enable a massive amount of sequencing data to be obtained, which also brings with it bioinformatics challenges and difficulties that are a bottleneck for many researchers (Scholz et al., 2012). In discussing this problem, Raes et al. (2007) have suggested that most comparative metagenomic approaches should consider at the outset of a proposed study such as the pitfalls and shortcomings of sequencing parameters. This includes all technical aspects, from sample collection, library preparation to sequencing and data interpretation (Raes et al., 2007). To avoid these pitfalls, they have proposed new standards of optimization known as MINIMESS for environmental shotgun sequencing projects (Raes et al., 2007). The MINIMESS is a proposed set of computational standard protocols to enhance the efficiency of pipeline analyses workflow for metagenomics data. The proposed MINIMESS consists of (1) basic sequence analysis, quantitation and qualitative – control analyses, (2) species richness estimation based on taxonomy analysis, (3) gene annotation, functional analyses, (4) species and gene coverage for *de novo* assembly, (5) evolutionary linkage between species, phylogeny analyses, as well as (6) addressing the common biological and technical measures such as GC content, genome size, read and contig lengths and complexity (Raes et al., 2007). Generally for deep-sequencing applications in

metagenomics, there are several key optimization goals. These include (1) sensitivity: detection of as many contigs and homologues as possible; (2) speed: comparison of run times – a few days (MiSeq) vs a few weeks (HiSeq); (3) accuracy: binning non-related sequencing data; (4) completeness: high quality assembled contigs, ideally whole genomes; and lastly (5) good bioinformatics tools and storage management for sequencing data (Deng, 2013)

1.3.3 Water Metagenomics

Marine and freshwater organisms are important components of aquatic ecosystems. From these environments, it is currently estimated that less than 1% of microorganisms are successfully recovered using traditional methods of culturing (Ghazanfar et al., 2010; Riesenfeld et al., 2004; Streit et al., 2004). The number of studies and investigations using NGS to characterise environmental samples has been growing steadily over the last few years. Numerous recent NGS studies have investigated aquatic microbial communinites and their diversity (Biddle et al., 2011; Bodaker et al., 2009; Breitbart et al., 2009; Djikeng et al., 2009; Fang et al., 2011; Morgan et al., 2010; Palenik et al., 2009; Venter, J. Craig et al., 2004; Woyke et al., 2009).

In general, recent freshwater metagenomic surveys have used one or a combination of three approaches 1) PCR and sequencing of 16S rRNA (rRNA/rDNA) hypervariable genes, 2) whole transcriptomics cDNA sequencing and 3) low coverage shotgun sequencing. The last method is the focus of this thesis. It provides for the profiling of several complex metagenomic samples at the same time with a resolution much greater than is possible with single gene sequencing and cloning approaches. However, it has not been the most widely used method of profiling, possibly due to its expense and the computational requirements for the fast matching of large numbers of NGS sequence reads to annotated sequences in reference databases. A review of some of the findings from the application of these different approaches in aquatic environments is given below.

In 2005, Cottrell and colleagues investigated bacterial communities using a metagenomics approach involving PCR enrichment of 16S rRNA genes and the fluorescence in-situ hybridization (FISH) technique. Fluorescence microscopy was used as an epidemiological tool to detect three types of bacterial groups in a river, *Actinobacteria, beta-proteobacteria* and *Cytophaga*-like bacteria. These were referenced against a dataset from extracted metagenomics samples for taxonomy allocation and analysis (Cottrell, Matthew T. et al.,

2005). The findings obtained from metagenomics and FISH techniques were different. For example the abundance of *beta-proteobacteria* were found to be underrepresented in the metagenomics libraries compared to FISH and vice-versa for *Cytophaga-like bacteria* (Cottrell, Matthew T. et al., 2005). In the taxonomic analysis, the microbial diversity profile found in the Delaware River from the PCR enrichment method (*Actinobacteria, Beta-proteobacteria and Cytophaga-like bacteria*) was different to that obtained from the traditional fluorescence tag (FISH) generated library. The authors concluded that the metagenomics technique offered significantly more coverage of the identified bacteria (Cottrell, Matthew T. et al., 2005). Metagenomics approaches have also been utilized as a tool for the evaluation and measurement of eco-toxicity levels within aquatic ecosystems. For example, Pope and colleagues (2008) investigated the cyanobacterial content in the phytoplankton (autotrophic 'plankton' community) of freshwater by constructing 16S rRNA gene libraries in bacterial artificial chromosomes (BAC) and using a combination of Sanger and pyrosequencing to describe the microbial community (Pope et al., 2008). A study by Venter and colleagues (2004), that was later followed up later by Rusch and colleagues (2007) demonstrated how NGS enabled the identification of genes and the elucidation of biogeochemical pathways from environmental samples collected from the Sargasso Sea to the Northwest Atlantic and Eastern Tropical Pacific oceans. The data generated using both Sanger and pyrosequencing, yielded more than 1.2 million unknown gene sequences from 1,800 identified unique genes along with 256 Mb of unique sequences (Venter, J. Craig et al., 2004). They also discovered 48 unknown bacteria phlylotypes which were closely related to the cyanobacterial genus *Prochlorococcus*. This experiment was extended by Rusch and colleagues in 2007 with their expedition called 'Sorcerer II' Global Ocean Sampling (GOS). Rusch and colleagues (2007) used metagenomic data to investigate the diversity, taxonomy, biogeochemical expression patterns and population genetics of the planktonic biofilm. During their expedition, forty-one different seawater samples were collected from a wide biodiversity zone of aquatic habitat during their 8000 km journey from the North Atlantic to the South Eastern Pacific ocean via the Panama Canal. In total, their result yielded approximately 7.7 million sequencing reads and about 6.3 billion base pairs from forty-one different environmental samples (Rusch et al., 2007). Metagenomics has also been used as a detection tool for aerobic and anaerobic microbes. This has enabled enhanced surveillance and risk assessment for transmissible diseases, especially for waterborne pathogens in drinking water (Breitbart et al., 2009; Djikeng et al., 2009).

In the future, it is predicted that freshwater eco-genomics research will become much more prominent compared to other metagenomic studies because of its importance in monitoring and maintaining the quality of drinking-water supplies in accordance with safety standards recommended by local city councils. There is a need to better understand more about environmental freshwater habitats as they can contain high levels of microorganisms derived from both soil and aquatic environments. Thus they have the potential to contribute significantly to the risk of pathogenic waterborne disease transmission (Cottrell, Matthew T. et al., 2005; Rusch et al., 2007; Sharma et al., 2009).

1.4 Overview of DNA Sequencing Technologies

DNA was firstly discovered and identified by Friedrich Miescher in the late 1800s and three-quarters of a century later, in the early 1950s, Watson and Crick together with Rosalind Franklin and Maurice Wilkins together proposed the arrangement of nucleotides known as the 'double helix' three-dimensional model of DNA structure. The chemical structure of nucleotides is made up of three important components: (1) nitrogen containing bases - the purines (Adenine, Uracil and Guanine) and pyrimidines (Tyrosine and Cytosine); (2) a five-carbon backbone sugar; and (3) a phosphate group (Pray, 2008).

In 1962, Frederick Sanger developed a technique for sequencing DNA, for which he was awarded a Nobel Prize. His method of dideoxy sequencing, which utilized radioactively labelled dideoxy (dNTPs) and the Klenow fragment of 'DNA polymerase I' provided a means to sequence short stretches of DNA of length less than 1000 base pairs. (Sanger et al., 1982; Sanger et al., 1977). The procedure involved synthesizing short strands of DNA in separate reactions which each was terminated by the random incorporation of one of the four ddNTPs. The terminated radioactively labelled sequences were then separated individually by polyacrylamide gel electrophoresis. The gel was then dried and exposed to x-ray film. The DNA sequence could be inferred by reading the order of dNTP terminations back from the gel front towards the gel wells (Sanger et al., 1977).

Next in 1980's, Sanger sequencing was modified and developed further with the invention of PCR amplification by Applied Biosystems, a commercial biotechnology company that introduced automated sequencers in 1987. Here instead of using radioactively labelled DNA, fluorescent labelled dideoxy dinucleotide tags (ddNTPs: ddATP, ddTTP, ddGTP and ddCTP) dye-terminators were used and each ddNTP emited a different fluorescent wavelength when

it passed through a laser beam in an electrophoresis system known as a 'capillary sequencer'. The differences in colour for each fluorescent labelled nucleotide was then captured by a CCD camera and recorded by a computer which processed the collected digital information an converted it into a DNA chromatogram. A disadvantage of this automated technique is that the cost of the DNA sequencing is relatively very expensive for large numbers of samples. With this method, the fragment length limit (850 to 1000 bp) is largely due to limited resolution of the acrylamide gel separation, although the formation of secondary structures and primer issues can also be a problem (Sanger et al., 1982; Swerdlow et al., 1990).

The human genome sequencing project was first proposed during the scientific meeting at Santa Cruz, California in May 1985 (Sinsheimer, 1989). Thereafter a special committee known as the International Consortium of Human Genome Project led by the United States and other genomic centres (i.e. UK Medical Research Council and the Wellcome Trust from Britain, European Genomics community, Ministry of Education, Science and Sports from Japan) was established to foresee the development of programmes to support the idea before being officially launched in early 1990 (Lander et al., 2001). With the introduction of the human genome project in the early 90's several improvements were made with the Sanger sequencing output and the most noticeable was the use of the shotgun cloning approach. Unlike in the conventional method of cloning the sequencing individual fragments, shotgun cloning involves large genomic DNA being broken randomly into smaller fragments by mechanical shearing. The fragments are then cloned, amplified, sequenced and assembled for data analysis (Myers et al., 2000). In 1995, Craig et al. (1995) used this technique to construct the genome of *Haemophilus influenzae* with 6x coverage on capillary sequencing technology (Fleischmann et al., 1995). Here copies of DNA (*H. influenzae*) were then cloned into a plasmid vector and sequenced from both directions using universal primers (M13-21) on automated sequencing machines before being assembled and analyzed for genome construction (Fleischmann et al., 1995). Later in 1998, the Celera Genomics Cooperation was founded and responsible for the sequencing of human genome and shortly after, they published the first genome sequence of fruit fly *Drosophila Melanogaster* with approximately 165 million base pairs of sequences that encoded ~13,600 genes via the 'whole genome shotgun' sequencing technique (Adams et al., 2000). This technique was utilised for the human genome sequencing project in 2001 by Celera Genomics. A total of 14.9 billion bp of DNA sequences, with approximately 2.91 billion bp from the euchromatin

regions (DNA, RNA and protein dense regions), were generated by using a whole-genome shotgun Sanger sequencing method from plasmid clones originating from five different individuals (Swerdlow et al., 1990). This faster whole genome shotgun Sanger sequencing approach has enabled the sequencing of a draft human genome to be completed within three years rather than thirteen years. In addition by adopting such a method, the overall project cost was reduced from 3 billion to 300 million dollars, with a cost reduction of approximately 90% compared to the previous method (Venter, J. C. et al., 2001; Venter, J. Craig et al., 2004).

A few years later at Stockholm in 1996, Mostafa Ronaghi and Pal Nyren developed an alternative method of capillary sequencing that utilized an emulsion bead-based system for PCR amplification and this technique is known as 'pyrosequencing' (Ronaghi et al., 1996). This new method is different from capillary sequencing. It detects the release of pyrophosphate (PP_i) during the synthesis of each nucleotide rather using fluorophore reversible dye terminators (Ronaghi, 2001). The first commercially available automation of pyrosequencing was the 454-Genome Sequencer FLX machine and this was released to the market in 2004 by Roche Life Science. In the 90's, Massive Parallel Signature Sequencing (MPSS) was introduced and this method resembled the current Illumina next-generation sequencing technique. MPSS is a probe/platform developed for the direct analyses of gene expression levels where individual mRNAs are counted via an adapter ligation technique (Ronaghi et al., 1996). A later development was the "sequencing-by-synthesis (SBS)" method, a proprietary technique developed by the company Solexa before being bought out by Illumina. This methodology led to the introduction of Illumina NGS platform known as the 'Genome Analyzer I' instrument in 2006 that utilized the 'sequencing by synthesis' method and dye-terminator technology. It is capable of generating massive numbers of parallel sequences simultaneously.

Recently, Life-Technologies (Applied Biosystems) introduced another different variant of second generation sequencing platform that utilises emulsion-based PCR and semi-conductor sequencing technology. Unlike other NGS technologies, the semi-conductor sequencing technique requires no moving and imaging parts. It relies on ionic (hydrogen ions) pH changes inside microwells which are detected by a pH meter. This platform is called the Ion-Torrent and was released to the market in 2011.

The current expansion of massive parallel sequencing technology either using pyrosequencing or reversible dye terminators has driven and broadened many metagenomics applications including microbial population taxonomic analysis. The parallel sequencing technique consists of several different methodology such as pyrosequencing (Roche 454), reversible dye terminator (Illumina) and newer Ion semiconductor sequencing (Ion-Torrent) (Metzker, 2010). These methods have all bought significant improvements to DNA sequencing and increased the data output, accuracy, efficiency and cost-effectiveness for metagenomics research.

In the study reported in this thesis, we evaluate data produced from protocols on two next-generation sequencing platforms, Illumina HiSeq or MiSeq and Life-Technologies Ion-Torrent PGM. These have been used to investigate a microbial community in an environmental sample taking from the Tamaki River. In our study, we use barcoded libraries which enabled multiple samples to be combined into a single lane of a flow cell. By utilising this methodology, we could pool more samples and still obtain high coverage for each sample, thus offering a very cost effective method for monitoring microbial activities within the aquatic ecosystem.

1.4.1 Illumina High-throughput Sequencing System

1.4.1.1 Illumina Sequencing Chemistry

For library construction, total genomic DNA is subjected to DNA fragmentation, end-repair, ligation, PCR amplification and a series of purification steps for the preparation of a DNA library (Bennett, 2004; Mardis, 2008b). For DNA fragmentation in the Illumina TruSeq protocol, long fragments of genomic DNA is sheared into 400 – 800 bp short fragments using either a nebulizer (fragmentation kit from Invitrogen) or a sonicator (Bioruptor) or by a hydroshearing instrument (Covaris). After fragmentation, the sticky 3’ overhanging nucleotide is repaired to create blunt end fragments via endonuclease activity (T4 DNA polymerase and Klenow fragment) prior to an adapter-ligation step (Mardis, 2008b). The product is ligated on both 3’ and 5’ ends, to adapters that are complementary to the oligonucleotide on the flow cell (Mardis, 2008a). For the sequencing of multiple samples, a unique identifier (an extra 6 base pairs) can be incorporated into one of the adapter sequences. With TruSeq protocol up to 12 unique indices can be run in a single lane on a flow cell for sequencing (Mardis, 2008b). After ligation, the product is purified and size selected to

lengths of 350 or 550 base pairs depending on the experimental design, prior to library enrichment via PCR and cluster amplifications.

Multiplex sequencing is gaining increasing interest particularly for metagenomics applications due to faster sample preparation protocols, high-quality data and cost effective methods for assessing larger numbers of environmental samples with great sequencing depths. For example, in one study, Zhou et al., (2011) evaluated the usage of Illumina Multiplexed Paired-end Sequencing Adapters (BIPES) on environmental samples that involved sequencing the 16S rRNA V6 region. In this study, the Illumina platform was chosen over 454-pyrosequencing because of the advantages of higher accuracy (short-read) and a larger amount of sequencing data available from a single run. In their experiment, a total of nine genomic DNA samples from the mangrove sediments were individually labelled using 16S V6 barcode-tags. This was achieved by amplifying the 16S V6 region via PCR and barcoded primers to produce 16S amplicons prior to Illumina sequencing (Zhou et al., 2011). The indexed samples were pooled into a single lane on a 2 x 100 bp HiSeq2000 run at the Beijing Genomics Institute (BGI). Two replicates included in this study produced very similar results with an overall error rate of 0.19% between both sequencing reads (Zhou et al., 2011). The authors concluded that the pooling of nine samples on a high-throughput sequencer using Illumina barcode-adapters was an approach that could significantly reduce cost while maintaining sequencing read quality, and thus was suitable for the quantification of microbial diversity (Zhou et al., 2011).

Another investigation led by Wong et al. (2013) highlighted the use of the Illumina TruSeq multiplexing technique (12 indexes) for DNA barcoding many organisms with small genomes and up to 96 samples in a single flow cell lane (ChIP-sequencing). ChIP-sequencing is a method combining chromatin immunoprecipitation (ChIP) with massive parallel sequencing to investigate the binding sites of associated DNA-proteins. To increase the efficiency of costing and sequencing data, the authors proposed a standardised barcoding preparation protocol for ChIP sequencing for DNA fragments of 100 bp to 500 bp long (Wong et al., 2013). The barcoded sequences were then built into the 3’ and 5’ ends of the Illumina adapter before it was ligated onto the DNA fragments (Wong et al., 2013). The authors’ findings demonstrated that such a technique was feasible (Wong et al., 2013) and can even be extended to accommodate larger numbers of samples (dependent on genome size

and coverage required) per sequencing run to reduce costs and the time required for sample preparation (Wong et al., 2013).

Recently, Illumina introduced a new type of library preparation protocol for NGS known as the Nextera and Nextera-XT preparation methods based on a high density *in vitro* transposition method. Nextera and Nextera-XT require only small amounts of starting material (50ng and 1ng respectively) for library construction. The enzymatic mixture (transposon-based method) greatly simplifies and increases the efficiency of DNA fragmentation and PCR enrichment in just under two hours, compared to the older complex and time-consuming TruSeq protocol which normally takes days for preparation (Lamble, S. et al., 2013). Numerous scientific publications have documented that sample preparation using both Nextera and Nextera-XT technology yields high quality libraries comparable to TruSeq generated DNA libraries with lower error rate and less bias (Caruccio, 2011; Lamble, S. et al., 2013; Lebreton et al., 2013; Perkins et al., 2013; Smith et al., 2012). Both protocols have been reported to be cost effective and capable of producing high quality sequencing data with small quantities of starting material and fast processing time.

1.4.1.2 Sequencing by synthesis (SBS) Illumina Sequencer

The chemistry behind sequencing by synthesis (SBS) was developed in the mid-1990s by Dr Shankar Balasubramanian and Dr David Klenerman at Cambridge University. In the summer of 1998, both of them decided to start-up a company known as Solexa which utilised the basics of reversible dye sequencing chemistry where a nucleotide is added one at a time with a pause in between for the imaging of single molecules (Davies, 2010). This sequencing technology was developed further by Illumina when the company bought Solexa in 2006 for $650 million leading to the introduction of Illumina GAII sequencing platform in 2009 (Davies, 2010). Today, SBS-based sequencing is the most preferred next generation sequencing technology and has been adopted worldwide.

The sequencing-by-synthesis (SBS) system utilises reversible dye-terminator technology for single nucleotide sequencing (Mardis, 2008a). After sample preparation, a genomic library of adapter-ligated DNA fragments is paired via fluidics capillary action with oligonucleotide anchors bound on the surface of the flowcell. The flowcell is a 1.4 mm wide enclosed transparency glass piece, similar to microscopic slide, grafted with 8 channels (Kozarewa et al., 2009; Mardis, 2008a; Stiller et al., 2009). Here adapter-ligated single-stranded DNA is isothermally cleaved, linearized, blocked, hybridized and amplified, forming a double-

stranded DNA bridge before being denatured again. The cycle is then repeated many times to generate millions of DNA clusters simultaneously (Holt et al., 2008). This process was performed in a separate instrument known as a cluster station. The newer MiSeq instrument comes with a built-in cluster station.

The sequencing process begins with the extension from the primer regions, of the DNA template, to produce the first cycle read. In each cycle, fluorescently labelled deoxynucleoside triphosphate (dNTPs) are competing and only one nucleotide is incorporated into the DNA template via DNA polymerisation and cleavage (Figure 1). The addition of each nucleotide, one at a time, is captured in real time by a light source (laser emission) and detected by the highly sensitive charge couple device (CCD) camera which records the release of fluorescent dye after DNA cleavage during DNA synthesis. The number of cycles determines the length of the sequencing read (Figure 1). To aid the intensity of the images during sequencing, a scanning-mix (HDP) and incorporation mix (ICB) are rinsed through the flowcell at each cycle, to enhance the intensity of the light emission (Linnarsson, 2010; Morozova et al., 2008).

In Illumina sequencing, the imaging technology is either two- or four- channel imaging. The four-channel imaging refers to the acquisition of four distinct image cycles for each nucleotide base (A,T,G,C) where each image is analyzed individually to determine the base calls before being combined together to form DNA sequences from each unique nucleotide cluster (Illumina, 2014) (Figure 1). Two-channel imaging is a newly developed imaging method that uses only two sets of dyes and images for each of the four nucleotides to determine the correct base of the sequences. Two-channel imaging uses only two laser excitation bands, the red and green filter bands, where clusters observed in red or green filter are translated as C and T bases respectively and clusters seen in both the red and green filters simultaneously are interpreted as an A base while unlabelled clusters are identified as a G base (Illumina, 2014) (Figure2).

In paired-end sequencing, after the completion of the first read, the DNA products are cleaved and washed away in preparation for the indexing reads. In the indexing cycle, the index 1 primer hybridizes with the DNA template and produces the first index read cycle, similar to the first read cycle, which is then washed away upon completion (Metzker, 2010) (Figure 3). Next the 3' end of the DNA template is de-blocked so that it will anneal with another free oligonucleotide on the flowcell. The index 2 primer will then start to read in the

same manner as index 1 before being washed away upon completion (Metzker, 2010) (Figure 3). To repeat the sequencing, DNA polymerase will then extend the second reverse strands of the DNA template forming the double stranded bridge before being linearized to a single strand where now the original forward complementary strands are washed away. The second read sequencing begins with the introduction of read 2 primer before being hybridised on the reverse strand and the cycle is repeated as in read 1 sequencing.

For preliminary data analysis, sequencing from the pooled images either from two- or four-channels imaging, are extracted and translated into a local sequence cluster database before being demultiplexed and separated based on the unique identification indices tagged during the sample preparation (adapter-ligation step) (Borgström et al., 2011). For each sample, similar sequences from Read 1 and Read 2 are clustered and assembled together to create multiple long contigs that overlap with each other for paired-end reads. This information can be used for further downstream bioinformatics to decipher, detect and translate digital information from paired-end read sequences to useful data information (deletions, insertions, inversions) (Figure 3).

Illumina Sequencing by Synthesis (SBS) Method via 4- and 2- Channel Imaging Cycle

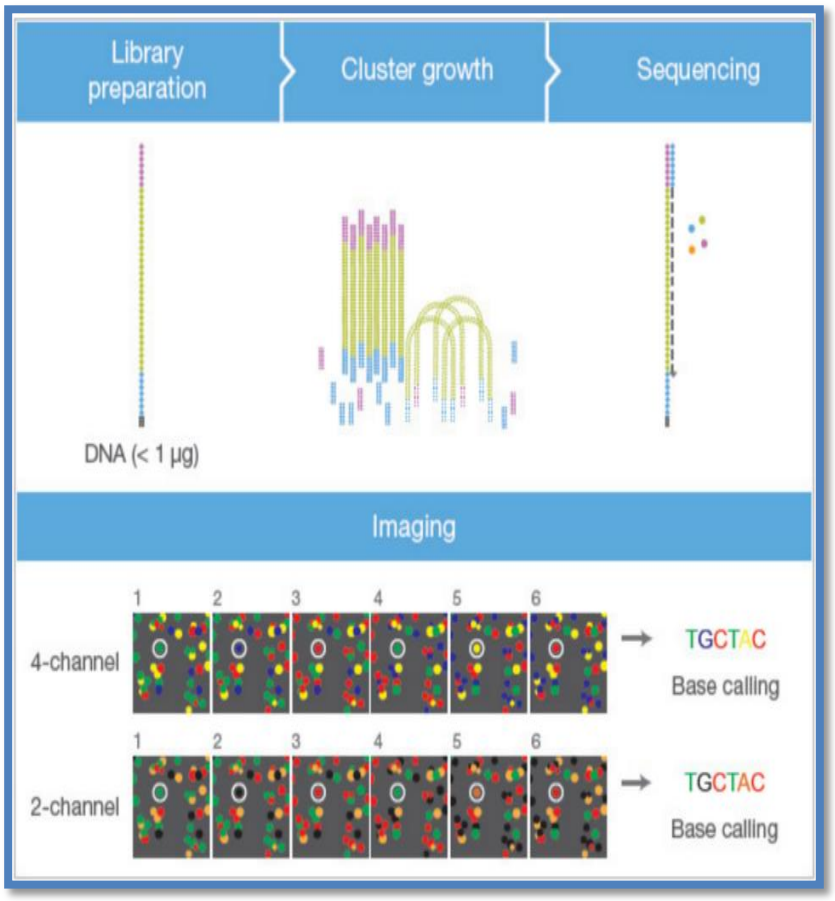


Figure 1 – Upper figure shows the sequencing-by-synthesis (SBS) workflow from sample preparation to sequencing meanwhile the bottom imaging show the base calling detection via 4- and 2-channel imaging detection technology using red and green laser filters. The latter has better accuracy and faster processing time (Figure provided by Illumina Inc).

Two-channel Imaging Technology

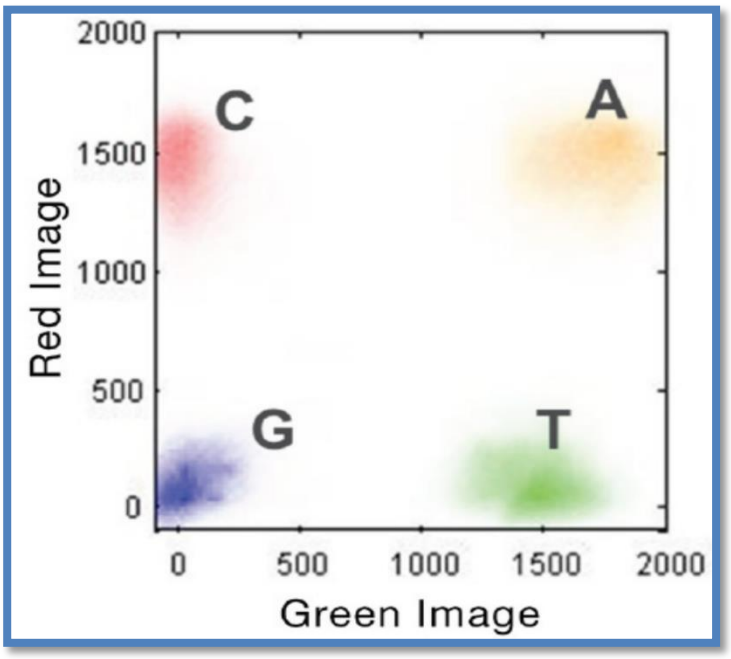


Figure 2 - In 2-channel imaging there are only two images captured (red and green filters) in determining the four nucleotides bases; the red colour represents the C base, the green colour represents the T base, meanwhile yellow (combination of green and red) represents the A base and lastly for the G base, it is blue-gray in colour with no specific filter colour coding (Figure provided by Illumina Inc).

Paired-end Sequencing Reads for Data Analysis

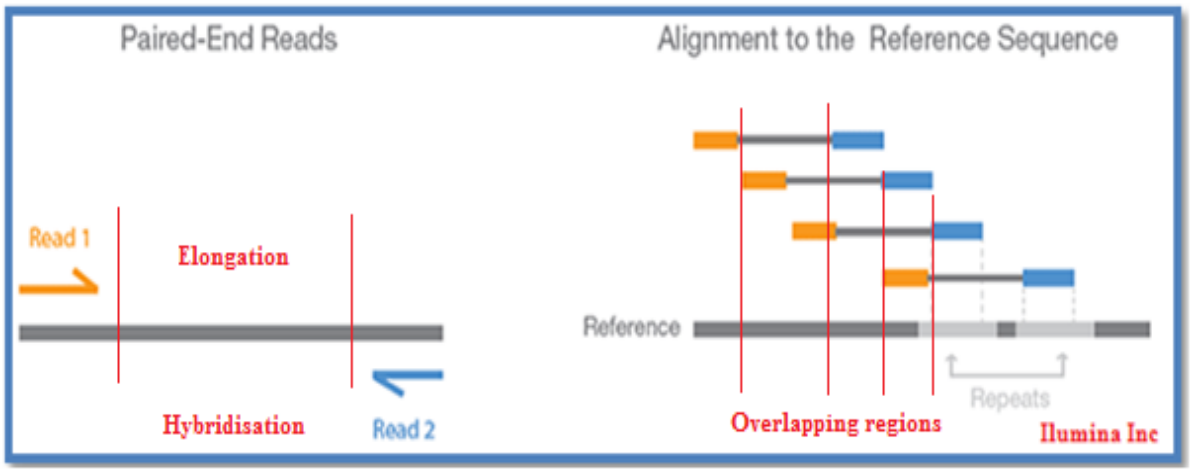


Figure 3 – Paired-end sequencing (left) showing Read 1 and Read 2 primers starting the elongation or extension of the DNA template after hybridization. The schematic on the right indicates how paired-end sequencing data can be used for elucidating the genome arrangement when aligned against a reference sequence. Paired-end sequencing can produce more accurate information due to the high number of overlapping regions of sequences and is particularly useful for difficult-to-sequence genome regions.

In February 2012, Illumina released the HiSeq 2000, a platform capable of producing 200 Gigabases (Gb) of sequencing data above a Phred quality score of 30 (Q score >30) for single and paired-end reads (25 Gb per day) with read length of 100 bp (two billion paired-end reads per run) in a dual flowcell configuration system (Ewing et al., 1998). At that time, HiSeq 2000 was the powerhouse for the sequencing industry with the highest sequencing output and fastest data generation rate with unprecedented simplicity and cost-effectiveness. For specifications and configurations, the HiSeq 2000 offered significant improvement over the GAIIx with the following upgrades (Figure 4): (1) dual camera system configurations for optimal 4-channel or colour dual surface imaging technology for capturing more clusters thus increasing its density and throughput, (2) dual flow cell technology giving flexibility to the researcher to choose different configurations in a single sequencing run e.g. first flow cell on 2 x 75 bp ChIP sequencing meanwhile second flow cell on 2 x 100 bp on whole genome sequencing, (3) high capacity reagent chiller for storage to ensure enough reagents for the entire sequencing run, (4) fully integrated fluidics for paired-end runs with a built-in reagent compartment for Read 2 re-synthesis and indexing run unlike the GAIIx that required a separate paired-end module, (5) in-built hardware and software interface control using touch screen monitor with separate keyboard, (6) real-time analysis software instead of separate pipeline analysis where the HiSeq 2000 monitors the sequencing run and provides quality metrics in real-time, (7) a Illumina cloud connectivity system (BaseSpace) where vital information from the instrument are pushed to an Illumina remote server for backup and flexibility to monitor the run anywhere and anytime via third party software (Illumina, 2013a).

The HiSeq 2000 sequencing system utilizes a separate ‘cBOT’ for the generation of clusters in a flowcell. Recently (2014/2015), Illumina released an upgraded version of its HiSeq sequencing system (HiSeq 3000/4000 and new instrument known as NextSeq 500) that have a much higher sequencing throughput with half the processing time of its predecessors, due to the latest 2-channel imaging technology and proprietary patterned flowcell with ‘nanowells’ enabling better accurate resolution of higher density flowcells.

Illumina HiSeq 2000 Instrument

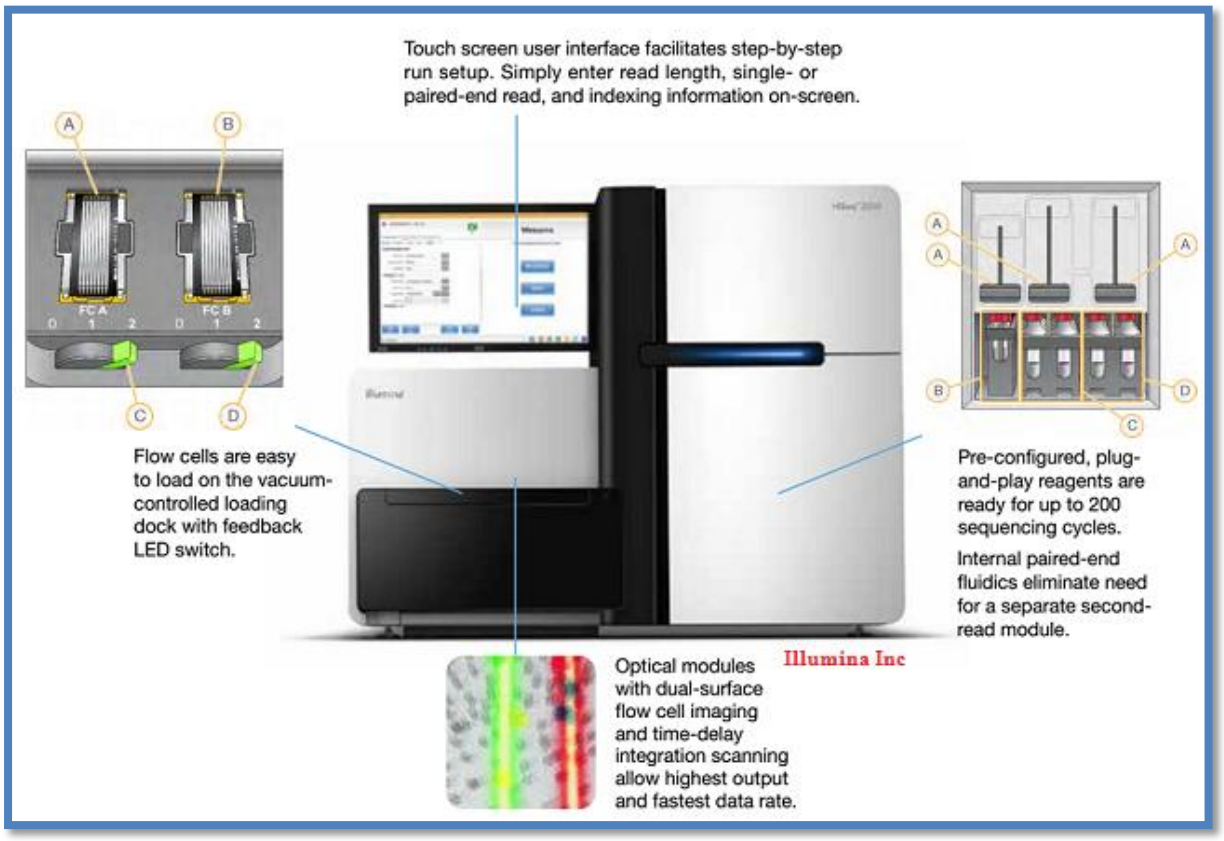
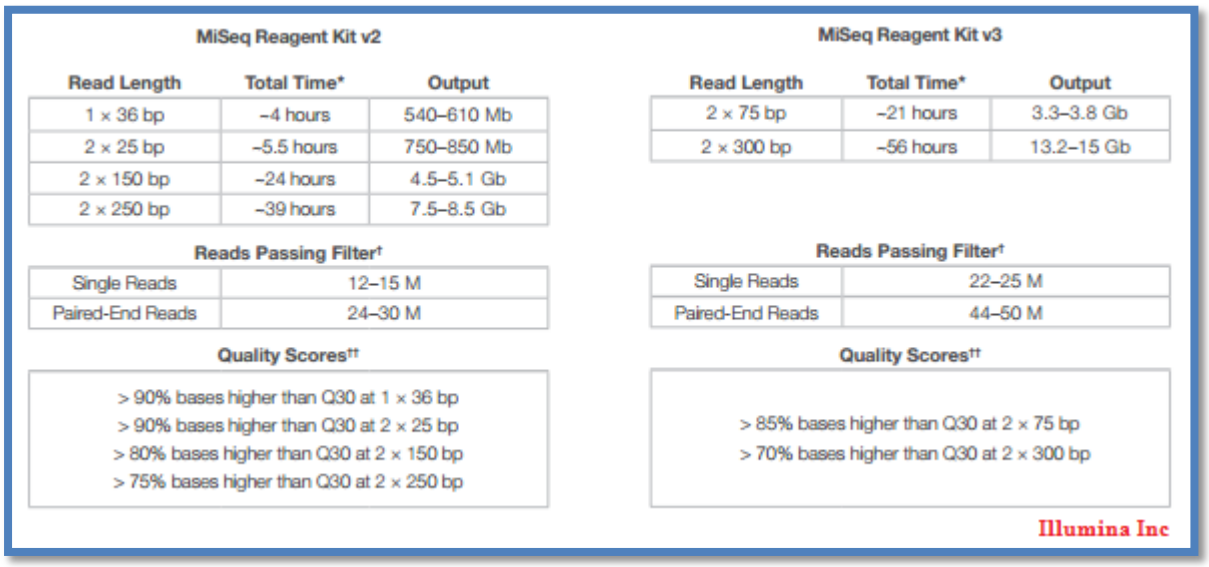


Figure 4 – A schematic overview of the HiSeq 2000 instrument showing the reagents compartment, optical modules with dual surface imaging technology and flow cell compartments that can hold two independent flow cells for a single sequencing run. All these improvements are now controlled by an integrated touch screen monitor with a simple intuitive interface.

The MiSeq desktop sequencer is a smaller sequencing system with much less sequencing data output per run compared to the high-throughput HiSeq sequencing system. Nonetheless, MiSeq are faster (data is obtained within hours) and are more economical to run (cheaper instruments and sequencing runs). The MiSeq sequencer is currently the only instrument that is capable of producing 2 x 300 bp paired-end reads (~25 million reads) in a single sequencing run. For sequencing, the MiSeq workflow offers wide selections for sequencing read length chemistry and caters for numerous different types of sequencing projects such as: genomic DNA sequencing, small-RNA and mRNA sequencing, CHiP sequencing, metagenomics sequencing, amplicon sequencing and others. Typically these types of smaller-scale projects need approximately 10-25 million reads with a sequencing output from 0.5 to 15 gigabases of data which is proving to be enough for comparing bacterial

communities based on 16S rRNA V3-V4 hypervariable region custom amplicons for most bacterial metabarcoding projects (Figure 5).

MiSeq System Sequencing Parameters



MiSeq Reagent Kit v2

Read Length	Total Time*	Output
1 × 36 bp	~4 hours	540–610 Mb
2 × 25 bp	~5.5 hours	750–850 Mb
2 × 150 bp	~24 hours	4.5–5.1 Gb
2 × 250 bp	~39 hours	7.5–8.5 Gb

Reads Passing Filter†

Single Reads	12–15 M
Paired-End Reads	24–30 M

Quality Scores††

> 90% bases higher than Q30 at 1 × 36 bp
> 90% bases higher than Q30 at 2 × 25 bp
> 80% bases higher than Q30 at 2 × 150 bp
> 75% bases higher than Q30 at 2 × 250 bp

MiSeq Reagent Kit v3

Read Length	Total Time*	Output
2 × 75 bp	~21 hours	3.3–3.8 Gb
2 × 300 bp	~56 hours	13.2–15 Gb

Reads Passing Filter†

Single Reads	22–25 M
Paired-End Reads	44–50 M

Quality Scores††

> 85% bases higher than Q30 at 2 × 75 bp
> 70% bases higher than Q30 at 2 × 300 bp

Illumina Inc

Figure 5 – Different sequencing chemistries available for various MiSeq sequencing projects. The projected output number of reads passing filter and quality scores are based on the Illumina internal sequencing PhiX control library with a cluster density of between 850 – 980 k/mm^2 using 2 x 250 bp version 2 chemistry, and between 1200 – 1400 k/mm^2 using 2 x 300 bp version 3 chemistry.

The MiSeq sequencing system has numerous advantages such as (1) a fully automated system from loading the flowcell, cluster generation, amplification and sequencing to data analysis in one instrument, (2) the reagents come in a cartridge without the need of advanced preparation apart from sample loading, (3) MiSeq software for system information, system configuration and monitoring of sequencing run, maintenance and data analysis (4) a one-way flow cell loading system with an auto correct clamping mechanism and integrated radio-frequency identification tag (RFID) for traceability, (5) an integrated fluidics system and lastly (6) real-time analysis (RTA) software providing valuable data analysis during sequencing. Recently Illumina released new MiSeq sequencing chemistry (version 3) for a 2 x 300 bp paired-end sequencing capable of generating fragment reads up to 550 bp with 25 million reads and 15 gigabases of sequencing data output.

1.4.1.3 Life-Technologies Ion Semiconductor Sequencing System

The Ion-Torrent sample preparation workflow involves more complex protocols in comparison to the Illumina workflow. Nevertheless, library construction is similar wherein 1 ng to 10 µg of starting genomic DNA material is subjected to numerous processing steps: (1) DNA fragmentation, (2) end-repair and 5'-phosphorylation, (3) adaptor-ligation and nick translation and (4) emulsion-PCR enrichment. For our study, a total of five different types of sample preparation kits were used to generate an Ion-Torrent library in preparation for sequencing: 1) Ion-Xpress Plus Fragment Kit, 2) Ion-Xpress Barcode Adapters, 3) Ion-Library Quantification kit, 4) Ion-PGM template generation (OT2) 400 bp kit and lastly 5) Ion-PGM sequencing 400 bp kit.

The Ion-Xpress fragmentation kit utilises enzymatic shearing for a greater and faster shearing process. The enzyme fragmentase used in this kit shears large genomic DNA > 10 kb into 100 - 800 bp fragments depending on the incubation time; normally about six to seven minutes which is similar to that for mechanical shearing. The enzyme fragmentase has two functions, first it randomly nicks dsDNA and secondly, it binds to the nicked site and cuts the opposite DNA strand thus producing a dsDNA overhang breakage known as a "sticky-end" which needs to be repaired as soon as possible to prevent re-annealing of both the complementary strands (Liu, Z., 2010). According to the manufacturer: Life-Technologies, enzymatic shearing and mechanical shearing produce similar results in terms of size distribution and sequence coverage. An advantage of using enzymatic shearing is that the fragment size can be controlled by diluting fragmentase and by using different incubation times tailored to generate the desired sized fragments for both AT- and GC-rich genomic libraries (Liu, Z., 2010). Following fragmentation, end-repair and phosphorylation, the enzymes Klenow-exo, and T4 DNA polymerase are used to repair the sticky ends of the dsDNA, prior to DNA ligation and emulsion-PCR enrichment.

Targeted-sequence enrichment serves several important purposes, (1) it increases the amount of prepared DNA template, (2) facilitates selection of molecules that are successfully adapter-ligated, (3) enables addition of indices for multiplexing technique and (4) enables incorporation of oligonucleotide sequences for the attachment of the library to the beads. There are two steps in the Ion-Torrent workflow where PCR enrichment occurs; the first is the amplification of the adapter-ligated DNA fragments and the second is the emulsion PCR.

Prior to sequencing the emulsified micro-reactors are broken apart to release the enriched DNA containing beads before being immobilized into the Ion-Chip for sequencing.

The Ion-Torrent Personal Genome Machine (PGM) was first introduced commercially in 2011. It sequences DNA by detecting the identity of the incorporated bases without the need for chemical luminescence dyes, with no requirement for camera optics, no light and no moving parts, hence making it simpler, faster and more affordable than other NGS platforms (Quail et al., 2012; Rothberg et al., 2011). Ion-Torrent uses semi-conductor chip technology that contains millions of tiny micro-wells under a huge sensing pixelated layer similar to the CMOS (complementary metal oxide semiconductor) light sensor chip found in the modern digital camera (Figure 6) (Quail et al., 2012). These tiny wells capture ionic pH changes during DNA sequencing which are then later translated to digital information. CMOS sensors are less sensitive than CCD sensors but are 10 ~ 100 times faster in processing light sources due to the ability to read and translate each pixel individually and simultaneously, producing excellent quality images with low background noises (Figure 7A). However in the Ion-Torrent PGM, the CMOS sensor has been modified and paired with an ISFET (Ion Sensitive Field Effect Transistor) sensor to sense chemical changes instead of changes in light (Figure 7B) (Rothberg et al., 2011). The sensor is positioned at the bottom layer over the electronics for transferring electrons during the transduction of voltage from the incorporation event (Figure 7B) (Rothberg et al., 2011) and is used as an independent pH monitor directly measuring the release of a hydrogen ion (H^+) during the incorporation of a nucleotide (Rothberg et al., 2011).

The semiconductor Ion-chip is a wafer-like square made from polycarbonate that contains millions of micro wells designed to hold and control fluidics on top of a CMOS and ISFET sensor arrays for the detection of electrical signals (Figure 7C) (Rothberg et al., 2011). These miniscule wells are designed to retain fluidics within the high conductivity material to ensure efficient electrical signal transduction (Figure 7D). Presently in the Ion-Torrent PGM sequencing system there are three types of Ion-chips (Ion-314, -316 and -318 v2) and each has a different sequencing output from 30 megabases to 2 gigabases with a total number of reads ranging from 400 to 5.5 million sequences (Shokralla et al., 2012). The exponential increase in the sequencing output was achieved by increasing the diameter of the semiconductor die cast area from an original size of 10.6 mm x 10.9 mm to 17.5 mm x 17.5 mm thus increasing the density output (Figure 7E and 7F) (Rothberg et al., 2011). However

the expansion of the area is limited by the CMOS sensor size and number of transistors. Further expansion would require a redesign of the Ion-chip system.

The sequencing process begins with the denaturation of the prepared DNA libraries inside the micro wells flooded with a dNTP solution. Within these wells, DNA nucleotides are incorporated one at a time via DNA polymerase. During this step whenever a nucleotide is incorporated into a single strand of the DNA, it releases a free hydrogen ion (H^+) as a by-product. The alteration in pH is then translated to a voltage signal before being recorded by the pH meter inside the microchip and translated later into digital information for each incorporated nucleotide base (Rothberg et al., 2011). Since each DNA nucleotide emits a different pH reading and voltage, the nucleotides can be confidently base-called individually without much error. In the event of identical nucleotide bases next to each other the voltage signal will give a double or more signal readout when base-calling, e.g. if three bases of thymine (T) are detected then the voltage signals for thymine will be increased to three-fold on a pH voltage meter. This sequencing process occurs across millions of wells in a microchip simultaneously which explains why the sequencing process only takes few hours instead of days for chemiluminescence detection. This semiconductor approach gives a read length range from 100 to 400 bp DNA fragments (Liu, L. et al., 2012; Quail et al., 2012; Rothberg et al., 2011).

For this project we have chosen both the Illumina MiSeq NGS platforms and Life-Technologies Ion-Torrent PGM as our preferred sequencing instruments. Both chain-termination based platforms and semi-conductor chip technology are equally capable of producing massive parallel sequencing sharing similarities in workflow, engineering configurations and sequencing chemistry. In the present work we compare the results from semiconductor sequencing with those obtained from Illumina SBS system for their performance while evaluating the data quality and the associative running cost.

Ion-Torrent PGM sequencing system

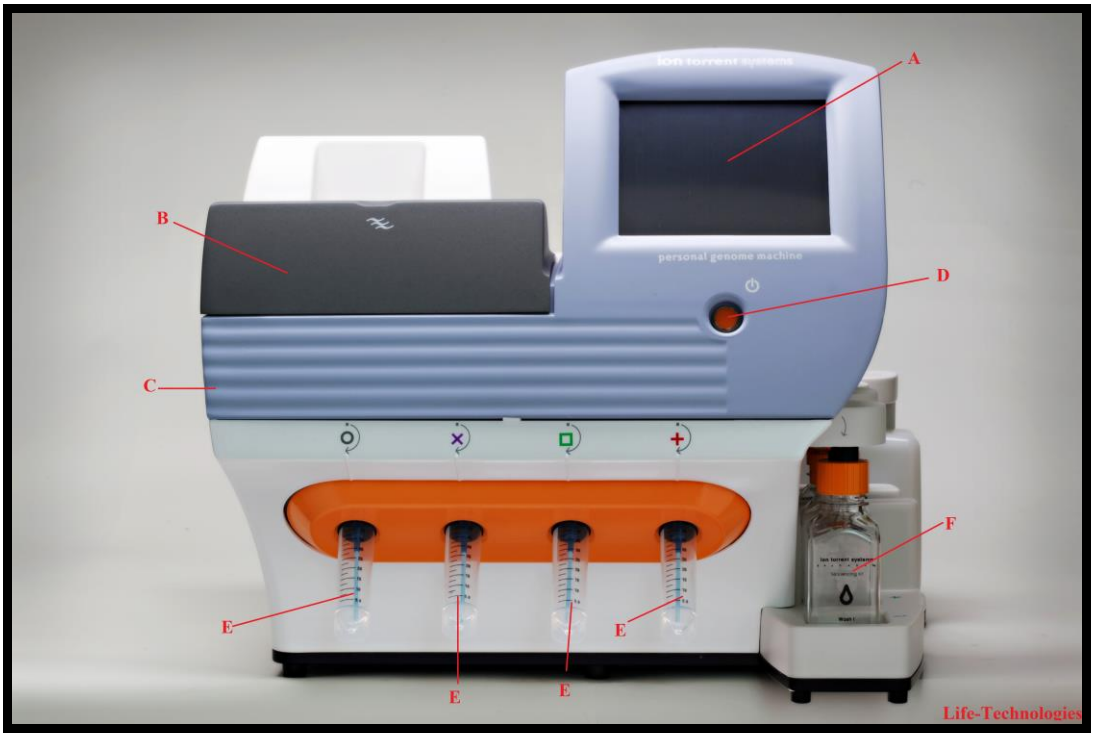


Figure 6 – Ion-Torrent PGM sequencer, A) touch screen control, B) Ion-chip loading deck clamping mechanism, C) special material grounding plate, D) power button, E) Reagent bottles, F) Wash bottles.

Ion-Torrent PGM sensor, well and chip design

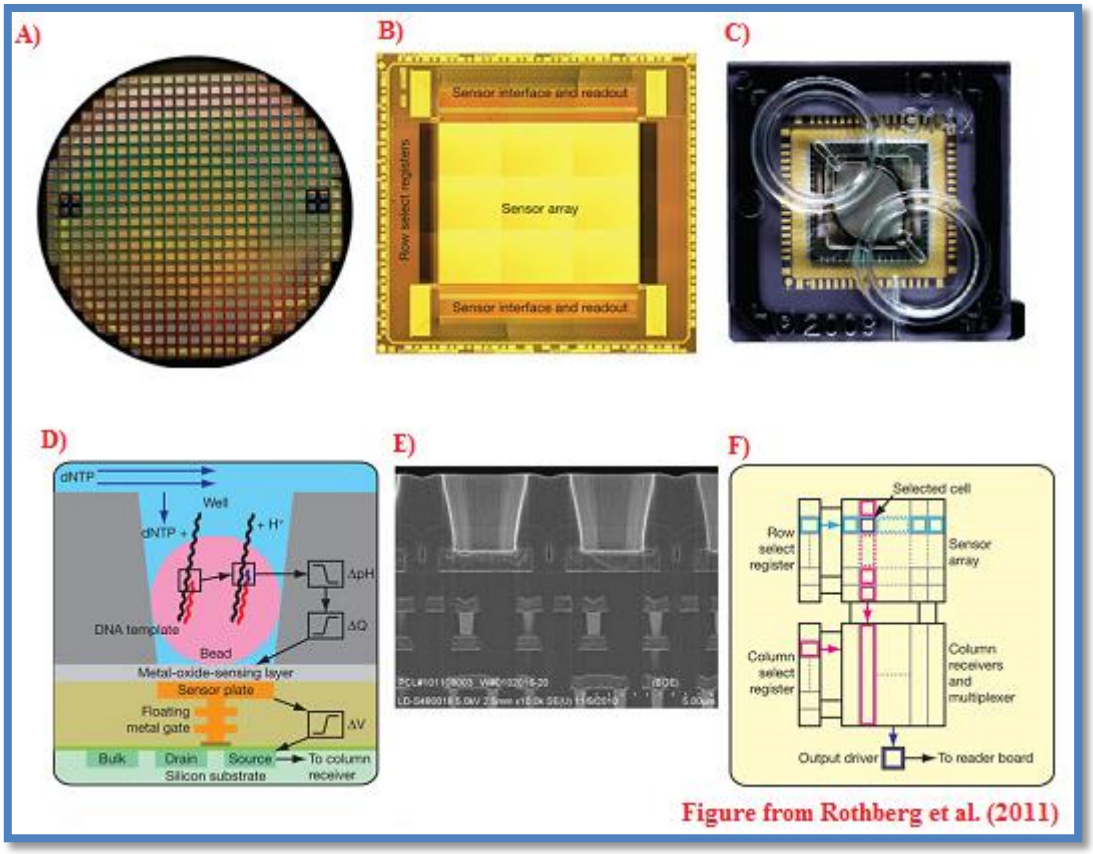


Figure 7 – Technology behind semiconductor sequencing, A) CMOS sensor build on a wafer shape polycarbonate die, B) underlying electronics and sensors board, C) upper surface of the Ion-chip showing location for addition of sequencing reagents, D) A schematic diagram showing the technology behind semiconductor sequencing with DNA template releasing H^+ ions which change the pH of the well - this signal is transformed into potential voltage and sensed by the under lying sensor and electronics, E) electron micrograph showing connection between miniscule well and ISFET sensor, F) schematic diagram for the sensor detection workflow in two-dimensional array.

1.5 Metagenomic analyses

Significant advances in computational genomics software programs used for assessing metagenomic data has been growing steadily since the introduction of NGS technology (Schleinitz, 2011). However the recent data explosion can be a roadblock for many researchers as there is difficulty in dealing with the amount of data generated from these high-throughput sequencing systems. Thus, new informatics toolboxes are needed in general but specifically, to investigate, evaluate and characterize the microbial communities within environmental samples (Gilbert et al., 2011). Statistical analysis of metagenomic data is confusing and time consuming as the management for such data information is complicated which requires a lot of effort in sorting and filtering the data before binning for taxonomy studies. These enormous datasets need to be properly sorted, aligned, assembled, quality-checked, trimmed, deciphered and viewed for proper interpretation to address various kinds of metagenomic computational needs (Chen et al., 2005; Schleinitz, 2011).

For most metagenomic analyses, a variety of bioinformatics tools and methods can be used to catalogue and describe microbial community profiles. However, no matter how complex the overall workflow is, a typical metagenomic computational workflow strategy is based on these steps 1) DNA sequencing, 2) Pre-QC checked and filtered metagenome data, 3) taxonomic analysis based on similarity-based classification (e.g. MEGAN5 or BLAST-based or clustering), 4) annotation for functional analyses (e.g. SEED and KEGG analyses) and in the case of mRNA studies 5) metatranscriptomics analysis (Kim et al., 2013). Currently, clustering techniques are typically used for the assembly of most metagenome data. A study conducted by Kelly and Slazberg in (2010) on "Clustering Metagenomic Sequences with Interpolated Markov models" presented two alignment tools for the characterisation of microbial communities (Kelley et al., 2010). Both tools utilised non-supervised sequence clustering with Interpolated Markov models (SCIMM) and supervised learning method PHYMM (PHYSCIMM) to identify clusters in K-means values (Kelley et al., 2010). The K-mean values here refers to the partition of the nearest cluster to the mean score (groups of similarities in sequences) where clusters are moved, binned and aligned before generating quality data points for classification purposes (Kelley et al., 2010). Both methods are highly accurate in predicting the maximum likelihood of cluster hits for large computational metagenomic data (Kelley et al., 2010).

The assessment of our freshwater metagenomic data was based on two fundamental approaches; the first was to determine what microbial taxa were present in the sample (qualitative assessment) and the second was to determine their abundance (quantitative assessment). The program known as 'MEGAN' (MetaGenome Analyzer) was used for this purpose. MEGAN is a tool for studying the properties of a metagenomic datasets and the clustering of taxa and their attributes using a phylogenetic methodology (Huson et al., 2007). Metagenomic data, i.e. sequence reads from environmental samples are compared to known reference sequences in a formatted database (e.g. a local version of a NCBI database) and the output is fed back into MEGAN for taxonomic classification and functional analysis (Huson et al., 2007).

BlastX was the original gold standard for matching DNA reads to a protein database. However, it is computationally too slow to be used for large scale metagenomics studies. A new analytical program known as PAUDA (Protein Alignment Using a DNA Aligner) became available in 2014 as an alternative alignment tool. PAUDA is a new method of blasting against the NCBI database where a large set of metagenomic data are converted first to protein sequences and then blasted against an index-built protein database for faster processing times due to simpler algorithm workflow. The PAUDA algorithm converts all DNA nucleotides (both database and data) first to protein sequences, to 'pseudo DNA' or 'pDNA', which is then modified into the 20 amino acid alphabet and further partitioned into four-clusters: [L, V, I, M, C], [A, G, S, T, P], [F,Y,W] and [E, D, N, Q, K, R, H] before being classified into a four-lettered protein amino acid alphabet A, C, G and T. All other characters are categorised as N. These converted pDNAs are then aligned and mapped against the pDNA database via Bowtie 2 (Huson, D. et al., 2014). Using this recoding algorithm significantly increases the performance and the processing speed up to 10,000 times faster than a typical BLASTing technique such as the 'blastx' and 'blastn' queries (Huson, D. et al., 2014). Also, PAUDA requires less than 80 CPU hours to analyse a dataset of 246 million DNA reads from permafrost soil compared to previous methods that required 800,000 CPU hours to reach the same functional analysis results (Huson, D. et al., 2014).

Using an alternative approach, Chen and Pachter (2005) demonstrated the use of other methods for data management and interpretation. Their study included tools such as the Lander–Waterman Model and Hidden Markov Models (HMM) for metagenome data (Chen et al., 2005). The Lander–Waterman theory is a mathematical equation used to calculate the

contigs ‘gap perspective’ of sequencing data (Sharon et al., 2009). Such methodology is normally used for the construction of a genetic map from functional analysis using COG and KEGG reference database based on an LCA-algorithm to screen for analogous gene-expression-level vectors (Chen et al., 2005). Here the information is then unfolded to reveal their genetic composition and used as a model to screen for genetic fingerprints from clusters of KEGG orthology (KO) accession number. Chen and Pachter showed that unfolding the genetic information prior to modelling is important for the construction of a fingerprint database that aids faster binning and characterization of microbial species (Chen et al., 2005). These recent advances in computational biology analyses, will help accelerate microbial biodiversity assessment and the search for key biological processes and specific biochemical pathways operating in microbial communities.

1.6 A role for metagenomics in studying freshwater environment.

Fresh water in one of the most recognized valuable natural assets of New Zealand. The rivers, streams and lake have great importance in sustaining New Zealand natural ecosystems. However, increasing levels of pollution due to human impacts means that increasing attention will be focused on the quality of our drinking water. As sequencing technologies are becoming more common and affordable, environmental profiling will be developed for the future studies of freshwater ecosystems. It will benefit from the use of different computational tools and multiple approaches to elucidate the nature of their diverse microorganism populations. Here, using metagenomics as a bio-monitoring tool provides an approach which will bring insight into our understanding of the diversity of biomes constituting freshwater environments. With metagenomics, direct measurement, both qualitative and quantitative can be used to compare both native and foreign biomes. Such comparisons will also help us to understand how microbial communities interact, integrate and diversify over time and under various environmental pressures.

1.7 Project Outline

The purpose of the work reported in this thesis has been to investigate high throughput sequencing technologies and protocols that are available for metagenomic studies of aquatic ecosystems. At the commencement of the current study relatively little was known about the advantages of different shotgun sequencing protocols for metagenomic applications. However, there are significant cost differences associated with different protocols. For example, the previous "gold standard" Illumina TruSeq library preparation protocol was 3-4x more expensive than the Epicentre (subsequently Illumina purchased) Nextera-XT library preparation protocol. At much greater cost, the TruSeq protocol also requires significantly higher labor costs and higher amounts of DNA sample. Thus an important motivation for the present work was a comparison and evaluation of next-generation sequencing (NGS) methodologies that might provide an alternative to Illumina TruSeq. This investigation included sampling and extraction techniques, different library preparation and sequencing chemistries, as well as data quality assessment. Two leading NGS sequencing platforms (Illumina MiSeq and Life Technologies Ion-Torrent) were compared. In this study, a single metagenomic sample was obtained by filtering water from the Tamaki River in Dannevirke. High quality DNA (evaluated by Nanodrop: OD260/280 $1.80 > x < 2.00$ and gel electrophoresis) was then used for library preparation and sequencing. DNA sequences were quality assessed using two bioinformatics pipelines SolexaQA and FastQC (Andrews, 2010; Cox et al., 2010) . Biological inferences in terms of taxonomic and functional profiles were made using PAUDA (Huson, D. et al., 2014) and MEGAN (Huson et al., 2007). This study identified a work flow involving enzymatic fragmentation as a time efficient and cost effective NGS protocol that has the potential to enhance investigation and understanding of complex processes in aquatic ecosystems.

2 Materials and Methods

2.1 Sampling Sites

For sample collection, we chose the Tamaki River nearby to Palmerston North. The Tamaki River (Figure 8a and 8b) is located on the edge of Dannevirke in the upper valley of the Manawatu River. This township is surrounded by the Ruahine Ranges, where the Tamaki River is broken into many smaller streams by numerous chains of mountains and hills (the Whangai, Waewaepa and Puketoi Ranges) (Wikipedia, 2004) . Dannevirke is a major dairy, beef cattle and sheep-farming town for the Tararua district in the Manawatu region of the North Island in New Zealand. The Tamaki River is used extensively by livestock farmers and industrial manufacturing plants (a wool processing factory, a concrete products factory and sawmill), with irrigation systems that are connected to the riverside. This site was selected because of serious declining trend of water quality and because sites in the area have recently been classified as high risk zones of water contamination. Recently, effluent comprising sewage and industrial waste which includes toxic chemicals, blood and milk have been deliberately drained into the river system. These activities have led to eutrophication, a condition where extra minerals/nutrients promote bacterial and algal growth which in turn increases the levels of nitrogen and organic compounds in the river water. Recent reports from the Ministry of Health have noted an increase in the number of cases of *Cryptosporidium* and *Giardia* protozoans in this region of the North Island.

Sampling Site Collection Map

Tamaki River

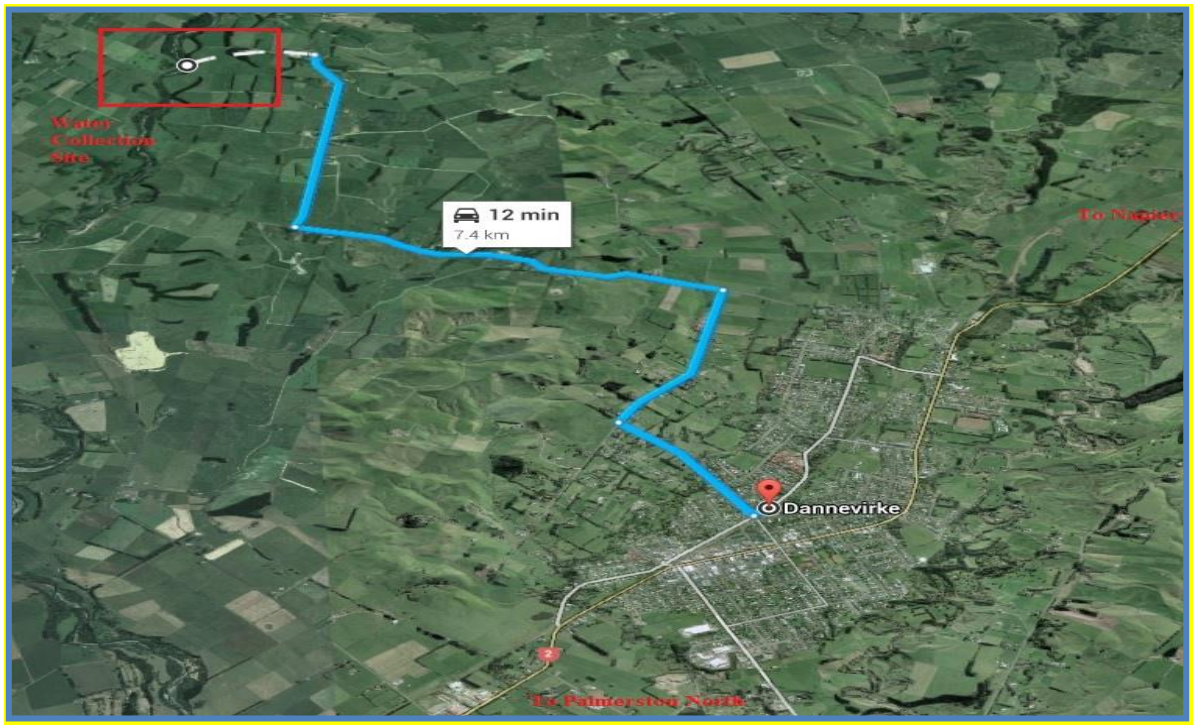


Figure 8a – Satellite image from Google Map showing the location of our water collection site on the Tamaki River near Dannevirke, Manawatu.

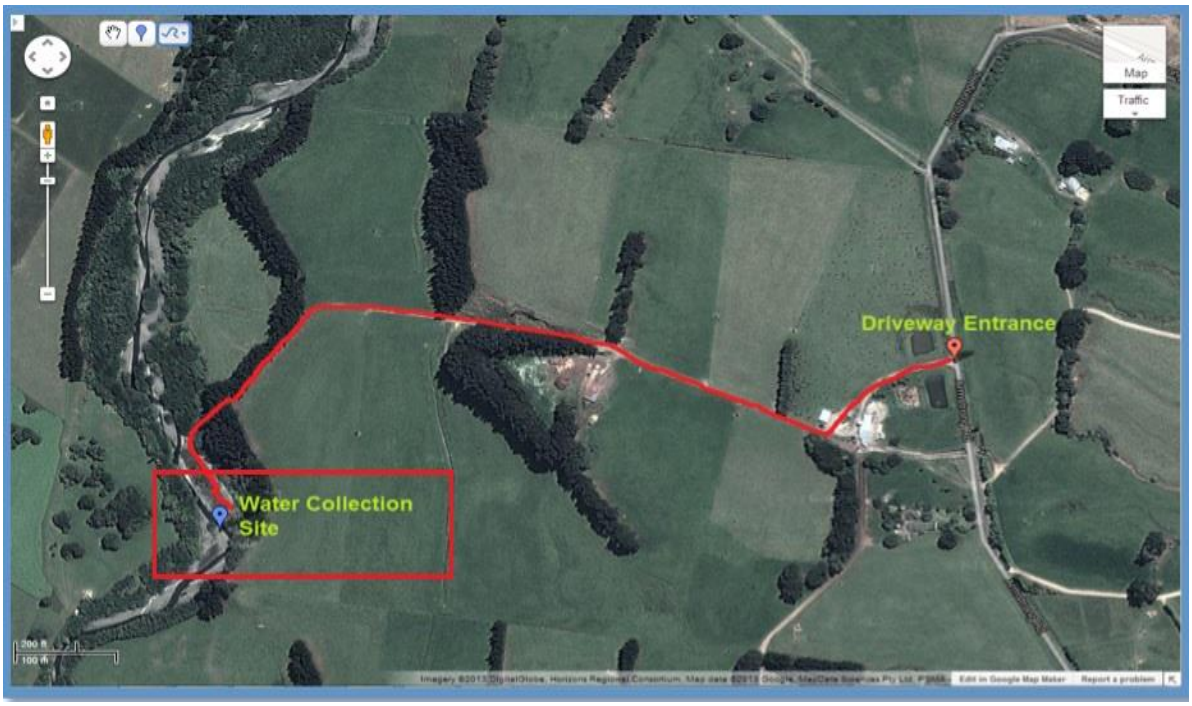


Figure 8b – Higher resolution satellite image from Google Map showing the GPS coordinates for the water collection site (40°09'43.2"S 176°03'50.5"E) and driveway entrance (40°09'38.1"S 176°03'50.5"E).

2.2 Sample Collection

Five 1 litre "grab" water samples were collected by myself from the Tamaki River (S 40 9.72, E 176 3.841). These were couriered to Massey University, Palmerston North in a chilly bin and washed and filtered through a special filtration apparatus vacuum kit purchased from Sigma Aldrich (Figure 9), usually no longer than 48 hours post collection. Filter paper was placed in between the filtration apparatus and the collecting flask and the flask was connected to a vacuum pump (Rocker Scientific Ltd, Model: Rocker 300 Oil-less vacuum pump, Cat No: 167300-22) to aid the filtration process. Two different pore size filters (Whatman, GE Healthcare) were used for the filtration (0.45 and 0.22 µm). Water samples were split into three batches of 100-ml to optimize the recovery capacity for each filter and thereby increase the concentration of DNA during the extraction process. The filtration apparatus was autoclaved between samples to prevent cross-contamination and filters were then stored at -80°C to prevent any degradation of microorganisms in the samples.

2.3 DNA Extraction

High molecular weight DNA (HMW) was extracted from the filter papers (0.45 and 0.22 µm) using the Metagenomic DNA Isolation Kit from Epicentre (Illumina Inc, Cat No: MGD08420). To remove the debris (filtrates) the filter paper was cut into quarters and placed into a 50-ml sterile falcon tube (Greiner Bio-one, Cat: 227 261). Next, 1 ml of filter wash buffer (10 mM Tris-HCl pH 7.5) and 2 µl (0.2% (v/v)) of Tween 20 (lauric acid ≥ 40% (w/v), myristic, palmitic and stearic acids) was added to wash off the microbes trapped on the filter paper. After washing and vortexing, the solution was pipetted into 1.7 ml microcentrifuge tubes (Axygen Inc, Cat No: 311-04-051) and centrifuged for 2 minutes to pellet the cells. Subsequently 2 µl of Ready-Lyse lysozyme [50% (v/v) glycerol solution containing 50 mM Tris-HCl (pH7.5), 0.1 M NaCl, 0.1 mM EDTA (pH8), 10 mM $CaCl_2$, 0.1% (v/v) TritonX-100, and 1 mM dithiothreitol)] and 1 µl RNase A [a 50% (v/v) glycerol solution containing 25 mM NaOAc (pH 4.6)] was incubated at 37°C for 30 minutes. Following incubation, 300 µl of 2X Meta-Lysis solution (20mM Tris, 2mM EDTA (pH 8), Tween 1% (v/v)) and 1 µl proteinase K (400 µg/ml) was added and the mixture was incubated again at 65°C for 15 minutes for cell lysis prior to precipitation. Next, 350 µl of MPC protein precipitation reagent (Epicentre proprietary) was added and followed by addition of isopropanol and two washes of 70% (v/v) ethanol before re-suspension in 50 µl of TE (10mM Tris, 1mM EDTA pH8.0) buffer. A total of 10 µg of DNA was extracted from the five 1 litre Tamaki River water grab

samples. An extract of DNA pooled from all grab samples was used to prepare libraries with the different metagenomics library protocols.

2.4 Colorimetric, microscopy and PCR tests

Coliform bacteria can be found naturally inside the intestinal tract of warm-blooded animals including humans. Thus, the presence of coliform bacteria normally indicates the water source is contaminated via leaking sewage or animal wastes. Hence for this reason it is often associated with other water-borne disease microorganisms such as G*iardia* and *Cryptosporidium*. Five water samples collected from the Tamaki River were subjected to preliminary colorimetric test at Massey University for direct measurement of organic and inorganic compounds of the coliforms. A colorimetric test will form a colour which distinctively changes upon detection of coliform bacteria. If a contaminated water source tests positive, a colour change will occur and vice versa if the test is negative. To further investigate the source of contamination, the water samples were screened with fluorescence microscopy (Nikon, IVABS, Massey University) with laser excitation and emission wavelengths at 490 and 530 nm under x40 and x100 magnification for *Giardia* and *Cryptosporidium* oocysts and spores (Sunnotel et al., 2006). Microscope slides were prepared, fixed and stained on glass slides by the Protozoa Research Unit (PRU, IVABS, Massey University) in accordance with guidelines for drinking-water quality management for New Zealand (Health, 2015; Sunnotel et al., 2006).

Next for molecular diagnosis, a PCR screening test for both *Giardia* and *Cryptosporidium,* was performed on the water samples by PRU (IVABS, Massey University) using specific target primers from the 18S rRNA gene and small-subunit (SSU) rRNA gene. For *Cryptosporidium* detection, the primers were as follows: AWA722F (AGTGCTTAAAGCAGGCAACTG), AWA1235R (CGTTAACGGAATTAACCAGAC), targeting the 18S rRNA gene, meanwhile for *Giardia* the primers were as follows: ABB97F (AGGGCTCCGGCATAACTTTCC), ABB220R (GTATCTGTGACCCGTCCGAG) (Rochelle et al., 1997). For the PCR amplification reaction mixture a 10 mM Tris-HCl (pH 8.3); 50 mM KCl; 0.01% gelatin; 0.2 to 0.5 mM each primer; 200 mM each dATP, dCTP, dGTP, and dUTP; and 2 U of DNA polymerase (Amplitaq; Perkin-Elmer Corp., Foster City, Calif.) in a 100 µl volume with 2 to 10 µl of template DNA (Rochelle et al., 1997). The reaction were performed using the following protocol: reaction mixtures were generally denatured at 94°C for 2 min, followed by 40 cycles of denaturation at 94°C, annealing for 1

min, and extension at 72°C for 2 min. A final extension incubation of 5 min at 72°C was included, followed by a 5 min incubation at 58°C to stop the reactions.

Metagenomics Water Filtration Process

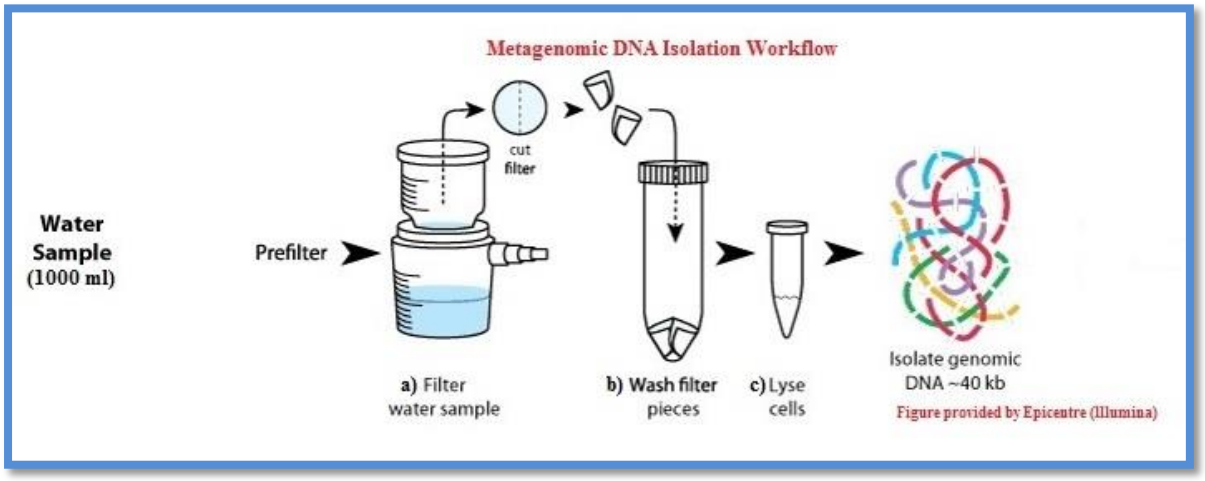


Figure 9 - One litre "grab" water samples were filtered through 0.22 and 0.44 µm filters. The microbes were then washed from the filters and their DNA extracted into a 20 µl volume of buffer prior to NGS library construction.

2.5 Pre-NGS library validation and quantification

Prior to sequencing of the collected 1 litre grab DNA samples, several steps were carried out to check the quality and quantity of the extracted genomic DNA (gDNA). To determine size and "intactness" of the gDNA, the samples were run on a 1% (w/v) agarose gel containing SYBr Safe, at 120V for 60 minutes in TAE buffer (40 mM Tris-Acetate, 2 mM EDTA; pH 8.0). An aliquot of 2 µl of each DNA sample was mixed with an equal amount of 2x gel loading buffer (0.2% (w/v) Bromophenol blue dye, 30% (v/v) glycerol in TE buffer) prior to loading for gel electrophoresis. DNA was visualised using a UV transilluminator (Bio-Rad Gel Doc 2000 UV system) and photographed. The camera aperture was set to auto-exposure of UV light for 44 seconds under high gamma settings (low gamma =0; high gamma = 255).

For quantification, two instruments were used: 1) a *NanoDrop* ND-*3000* spectrophotometer (Thermo Scientific) and 2) a *Qubit* 1.0 Fluorometer (Life Technologies). For the *NanoDrop* ND-*3000* spectrophotometer, 2 µl of sample was used to measure the concentration and purity of the gDNA samples. For DNA purity, the ratio of the absorbance readings at 260nm and 280nm should fall between 1.8 and 2.0. Next, for the Qubit quantification, 2 µl of purified DNA was mixed with 198 µl of Qubit working solution. This contains a fluorescent

dye that binds quantitatively to DNA, RNA and protein. Standards for DNA, RNA and protein were prepared for instrument calibration and used to plot size-standard curves.

2.6 Construction of NGS Metagenomics libraries

DNA libraries were prepared using four library preparation kits, the Illumina Nextera and Nextera-XT, NEXTFlex PCR-free (Bio Scientific) and lastly Ion-Xpress 400 bp (Life-Technologies) kit.

2.6.1 Preparation of libraries for Illumina sequencing

2.6.1.1 Nextera and Nextera-XT DNA Sample Preparation Method

"Nextera" and "Nextera-XT" (Illumina Inc, Cat No: FC-121-1230, FC-131-1024) protocols differ from other NGS protocols in that they use less starting material for the preparation of a NGS library. They require 100 ng and 1 ng of starting materials respectively. Both protocols use an enzymatic shearing method via the enzymatic *transposase* action (refer to appendix for Nextera *transposase* sequences). Prior to library construction, DNA was quantified using the Qubit 1.0 fluorometer and three different types of assays (DNA, RNA and protein) (Life-Technologies, Cat No: Q32866). To begin the fragmentation process and simultaneously tag the fragments, 25 µl of tagmentation buffer and 5 µl of tagmentation enzyme (Illumina Inc) were added to the quantified DNA and incubated at 55°C in a thermocycler (GeneAmp 9700, Life-Technologies, Cat No: N8050200) for 8 minutes (Figure 10). Immediately after incubation the fragmentation process was stopped with the addition of 5 µl of stop fragmentation buffer (NTA) (Illumina Inc) and incubated for 5 minutes at room temperature. The fragments were then cleaned and purified using a Zymo Clean and Concentrator-25 column (Zymo-Research, Cat No: D4006) as per the manufacturer's instructions and eluted in 20 µl of Illumina resuspension buffer (Illumina Inc).

For PCR enrichment, 20 µl of the purified fragments from the column eluates were mixed with 25 µl of Nextera PCR master mix (Illumina Inc) and 5 µl of PCR primer cocktail (Illumina Inc) along with 1 µl of each allocated index 1 (i7) and index 2 (i5) primers (refer to appendix for Nextera 96-index sequences) (Figure 10). Cycling conditions for the PCR reaction were as follows: 72°C for 3 minutes, 95°C for 30 seconds, 15 cycles of 95°C for 10 seconds, 55°C for 30 seconds and 72°C for 30 seconds, 72°C for 5 minutes and lastly, a hold at 10°C. Next the enriched fragments were purified using 30 µl of AMPure XP beads (Beckman Coulter Inc, Cat No: A63881) as described previously, washed twice with 80%

(v/v) ethanol and eluted in 30 µl of Illumina resuspension buffer (Illumina Inc). The size range and quality of the library was measured on both a Bioanalyzer 2100 instrument (Agilent Technologies, Cat No: G2938-68700) and a Qubit 1.0 fluorometer (Life-Technologies, Cat No: Q32866). Next the prepared library was diluted to 2 nM in pooling buffer (10 mM Tris-HCL pH 8.0, dH_2O and 0.1% (v/v) of Tween-20), prior to NGS library cluster generation and sequencing on a MiSeq instrument.

Nextera DNA Library Preparation

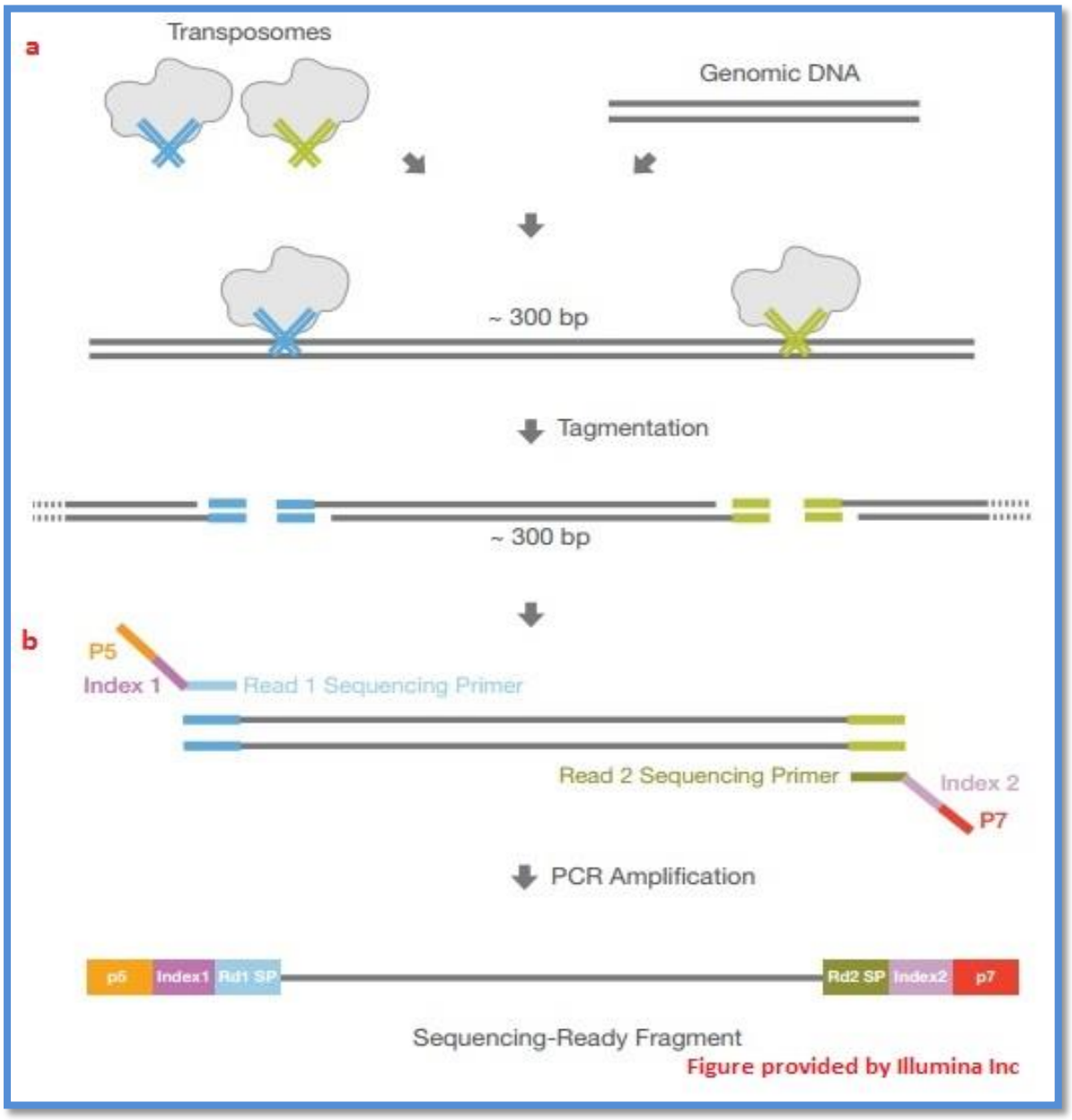


Figure 10 – (a) Nextera sample preparation uses a 'transposase' enzyme to fragment and tag DNA in a single step. (b) Primer adapters for read 1 and 2, along with individually bar-coded index i7 and i5, are added for PCR amplification before sequencing.

2.6.1.2 NEXTFlex PCR-free (Illumina Compatible)

DNA from the Tamaki River water sample was subjected to a 'low-throughput library preparation' protocol (Bioo-scientific NEXTFlex DNA PCR-free Sample Preparation Guide, Cat No: 514110, July 2011) prior to DNA sequencing. The DNA library was prepared by the author (as Senior Sequencing Technician) at the Massey Genome Service, Massey University, Palmerston North. To begin the mechanical shearing process, the nebulizer was attached to a small vinyl tube and 50 µl of DNA (2µg) was mixed with 700 µl of nebulisation buffer (50 % (v/v) glycerol and 50 % (v/v) TE Buffer in the nebulizer (Life Technologies; Cat No: K7025-05). The mixture was chilled on ice for 5 minutes, then connected to a compressed air source (Nitrogen, BOC, Cat No: P1 152 28) and nebulised at 35 PSI for 6 - 8 minutes. The shearing process takes place as the nitrogen gas passes through the solution in the nebulizer. Immediately after fragmentation, the sheared DNA was purified through a ZYMO PCR purification column (Zymo Research, Cat No: D4004)) by mixing with 2 ml of Zymo DNA binding buffer *(*4.2 M guanidine hydrochloride and 40% (v/v) isopropanol,) passing the mixture through a column, washing twice with 500 µl of Zymo washing buffer *(*1.0 M NaCL, 50mM MOPS, pH7.0, 15% (v/v) isopropanol,) and eluting in 16 µl of elution buffer *(*10 mM Tris-Cl, pH 8.5).

To monitor the DNA size distribution after fragmentation, the sheared products were analyzed in a Bioanalyzer (Agilent Technologies, DNA 1000 kit, Cat No: *G2940CA*). To start the analysis, the Bioanalyzer priming station (Agilent Technologies, Cat No: G2938-68700) (Figure 11) was set up according to the manufacturer's protocols by aligning the syringe cap and base plate to the correct position. To prepare the gel matrix, 25 µl of dye concentrate was aliquoted into the vial containing DNA gel. This was vortexed and then transferred to a spin-column and in which it was spun for 15 minutes at 6,000 rpm. For gel loading, 9 µl of gel-dye mix was aliquoted into the well-marked 'G' on the chip along with 5 µl of marker in the remaining wells (Figure 11). 1 µl of DNA sample was then aliquoted into the electrophoresis gel wells. 1 µl of concentrated DNA ladder was added to a separate well. The Bioanalyzer chip (Figure 11) was then vortexed at 2,400 rpm for one minute before loading onto the Bioanalyzer 2100 instrument.

Agilent Bioanalyzer 2100 DNA Chip

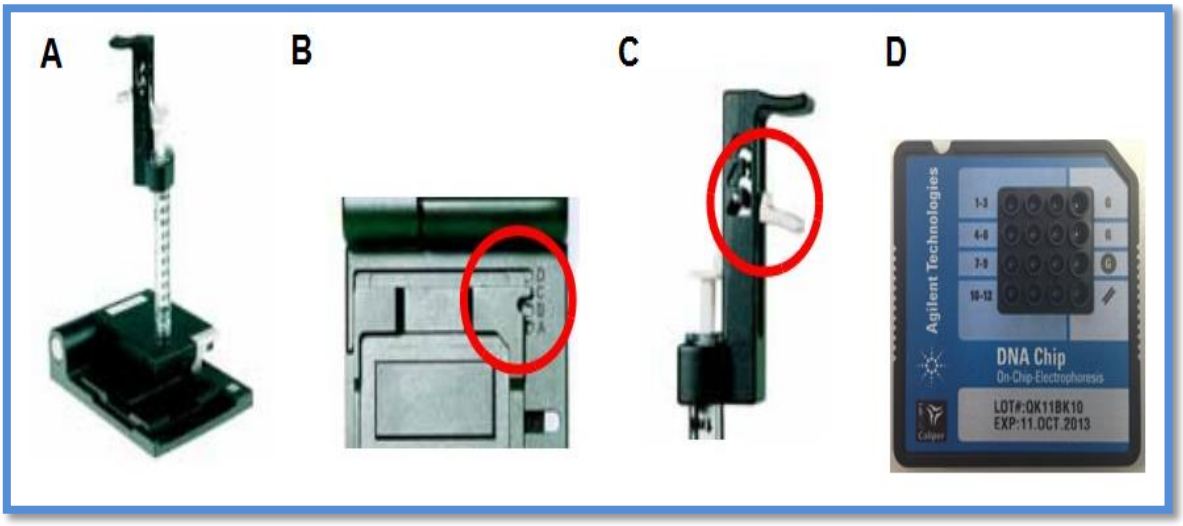


Figure 11 - A) Agilent bioanalyzer priming station, B) priming station base plate aligning to a correct position C) syringe lock clip was set to lowest position D) the DNA 1000 chip showing the position of the wells.

Once it was determined that the size of the DNA fragments was correct, the DNA was repaired by replacing the 5' overhangs with a phosphate group to allow the ligation of the adapter and by removing the existing 3' overhangs to create blunt ends. For end-repair, 40 µl of fragmented DNA was mixed together with 7 µl of end-repair Buffer Mix (BIOO-Scientific, NEXTFlex Cat No: 5140-05) and 3 µl of End-Repair enzyme mix (BIOO-Scientific, NEXTFlex Cat No: 5142-02) and incubated in a thermocycler for 30 minutes at 22°C (Figure 12). The PCR-free protocol used an alternative method of library size selection involving the Solid Phase Reversible Immobilization (SPRI) bead system. The SPRI system uses magnetic particles (beads) coated with charged carboxyl groups that bind reversibly to DNA in the presence of polyethylene glycol (PEG) and salt (usually NaCl).

In the PCR-free protocol, the DNA was purified by collecting the beads on a magnet following successive washes and elution in a low salt buffer. The size selection process is dependent on the relative concentration of beads to DNA, where the lower the ratio, the larger will be the DNA fragments. For size selection, 160 µl of magnetic beads (Beckman Coulter Inc, Cat No: A63880) was added to the sheared DNA and washed twice with 200µl of 80% Ethanol (EtOH) before being eluted in 15 µl of elution buffer (Figure 12). Next 3.5 µl of adenylation Mix (BIOO-Scientific, NEXTFlex Cat No: 5142-02) was added into 17 µl of

eluted DNA and incubated on a thermocycler for 30 minutes at 37°C (Figure 12). Following A-tailing, 31.5 µl of Ligation Mix (BIOO-Scientific, NEXTFlex Cat No: 5142-02) and 1.5 µl of DNA Index Adapter 2 (BIOO-Scientific, NEXTFlex Cat No: 514101- for Illumina compatible adapter sequences please refer to the appendix section) were added into 20.5 µl of adenylated DNA and incubated again for 15 minutes in thermocycler at 22°C prior to a post-ligation clean-up using AMPure XP beads (Figure 12). This method eliminates PCR enrichment and size selection steps which normally require an agarose gel. The fragments were quantified and checked for size distribution as before, using the BioAnalyser and the Qubit fluorometer. Prior to sequencing, the library was diluted to 9 pM which was the optimal loading concentration according to the manufacturer's protocol (Illumina MiSeq System User Guide, Part No: 5027617 Rev. F, Nov 2012). Using the NEXTFlex protocol, the Tamaki River gDNA was successfully fragmented, end-repaired, and size-selected for 550 bp fragments.

NEXTFlex PCR-free Library Construction Steps

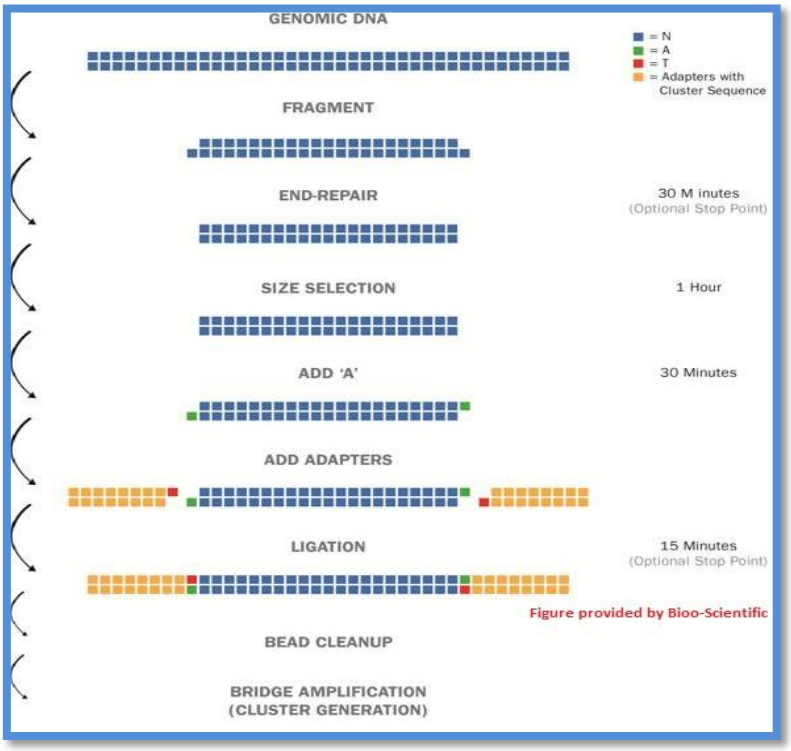


Figure 12 – With the NEXTFlex protocol, 1-3 µg of starting material (gDNA) is sheared to smaller fragments. The end-repair process and size selection process are merged into a single step via the SPRI beads system that binds to the DNA accordingly to the concentration of magnetic beads. After adenylation, a new enzymatic mix is employed to enhance the adapter ligation step prior to cluster generation, without the need for a PCR enrichment step.

2.6.1.3 Ion-Torrent Library Preparation

Approximately 0.5 µg of extracted DNA from the Tamaki River was sent to New Zealand Genomics Limited (NZGL, Auckland). The DNA was quantified on a Qubit 1.0 fluorometer prior to the delivery of samples. Next the samples were fragmented, ligated and enriched using an emulsion PCR-based amplification method carried out by the Centre for Genomics, Proteomics and Metabolomics (CGPM, NZGL), University of Auckland (Figure 13).

Ion PGM Sequencing System Workflow

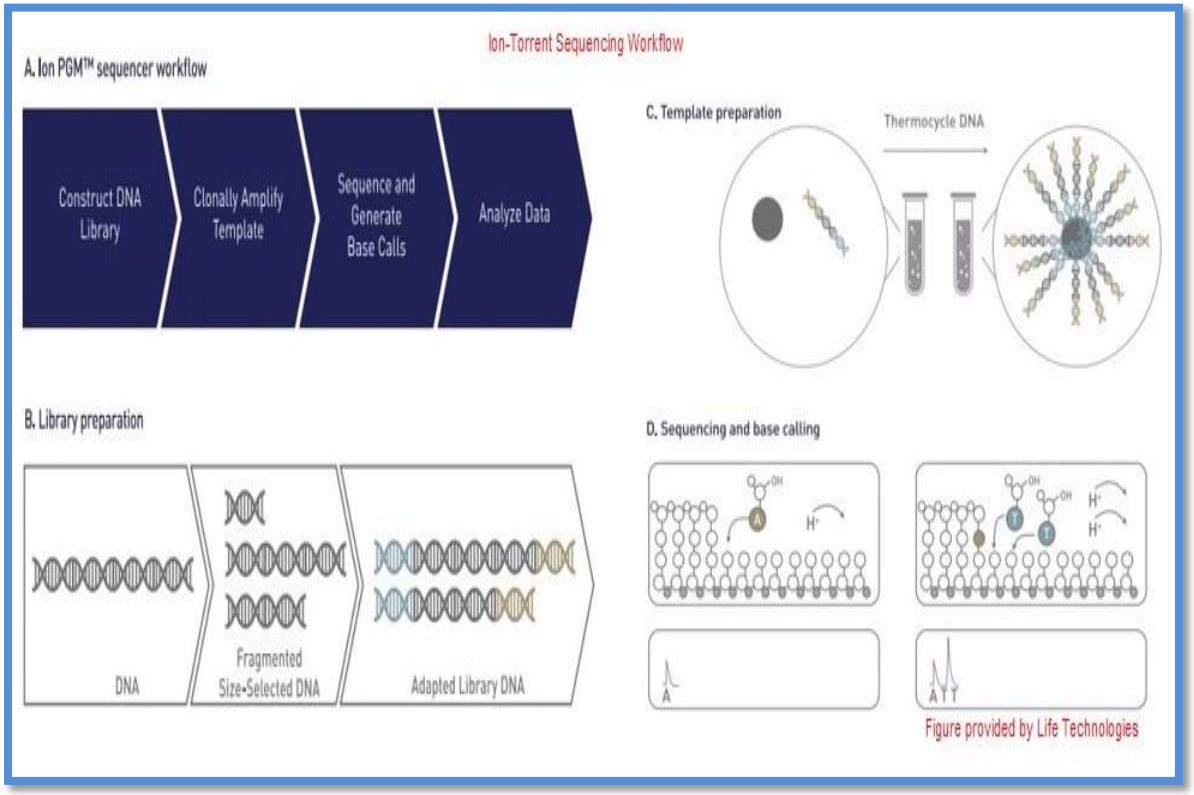


Figure 13 – Ion-Torrent sequencing can be achieved within hours due to the speed of semiconductor ion sequencing (A). Genomic DNA is fragmented and size selected using a SPRI bead system, before the adaptor ligation step (B). Next the adapter-ligated DNA is bound to Ion Sphere particles and amplified (C). These products are then loaded onto an Ion Chip and sequenced on the Ion-Torrent machine (D).

2.6.2 Illumina Sequencing

2.6.2.1 MiSeq Sequencing System

Here the metagenomic library from Tamaki River was quantified and quality checked via the Qubit fluorometer and the Agilent Bioanalyzer (using DNA 1000 Chip) as previously described, prior to dilution in resuspension buffer (10 mM Tris-HCl, 10% Tween-20) to 2 nM. Sequencing on the MiSeq instrument (Illumina Inc, Cat No: SY-410-1003) was carried out at the Massey Genome Service (MGS, Massey University) by the author. For sequencing, reagents (MiSeq Reagent Kits v2, 300 cycles, Cat No: MS-102-2002) were thawed in a water bath no more than 30°C for 1 hour. While the reagents were thawing, 10 µl of the library was mixed with 10 µl of 0.2N sodium hydroxide and incubated for 5 minutes at room temperature for denaturation. Next, 20 µl of denatured DNA was diluted to 10.5 pM (dilution with pre-chilled HT1 (hybridisation) buffer. MiSeq System User Guide, Illumina Inc, Cat No: 15027617). Following dilution, for metagenomics sequencing 25% (v/v) of 12.5 pM PhiX control library (Illumina Inc, PhiX Control v3, Cat No: FC-110-3001) was spiked into the denatured DNA. Thereafter, 600 µl of sample library was loaded into the reservoir on the MiSeq reagent cartridge. For clustering, both the reagent cartridge and the flow cell were loaded into the MiSeq instrument and the sequencing on the MiSeq was performed according to the manufacturer's protocols (Figure 14). The sequencing process was subjected to real time matrix monitoring where quality statistics such as data intensity, cluster density, Q-scores and % of errors were tabulated within a sequencing analysis viewer (SAV) software. After sequencing, the data was filtered, trimmed (removal of Illumina adapter) and pooled together to generate a 'fastq' format file for computational analysis.

Illumina MiSeq Sequencing Preparation

Figure 14 - Illumina MiSeq instrument is the only 'all in one' sequencer capable of producing clusters and sequencing under 'one roof'. Sample preparation and automated real time data analysis required less than a day for sequencing 2 x 150bp paired-end reads.

2.6.2.2 Ion-Torrent Sequencing

The sequencing was performed on a Personal Genome Machine platform (Ion-Torrent PGM, Life Technologies, Cat No: 4462921) at the Centre for Genomics, Proteomics and Metabolomics (CGPM, NZGL), University of Auckland (Figure 15). According to the sequencing provider, fragments of 200 – 250 bp were selected and processed using the Ion-Torrent 318 chip with a 65 cycle sequencing kit, that is capable of generating 2.0 Gb of sequence data with a minimum of 4 –5.5 million reads per run (Figure 15). The concentration of the library was at 9.5 pM for loading, as per the NZGL Ion-Torrent protocol. Sequencing was performed using the Ion Sequencing 200 bp Kit (Life Technologies, Cat No: 4474004) following the manufacturer's protocols. After sequencing, the data was processed using the computational software 'Torrent-Suite (v1.5)' for preliminary bioinformatics analyses, with a default output format in 'fastq' instead of SSF file format. The data was downloaded via a temporary FTP website generated by NZGL, Auckland.

Ion-Torrent Sequencing Chemistry workflow

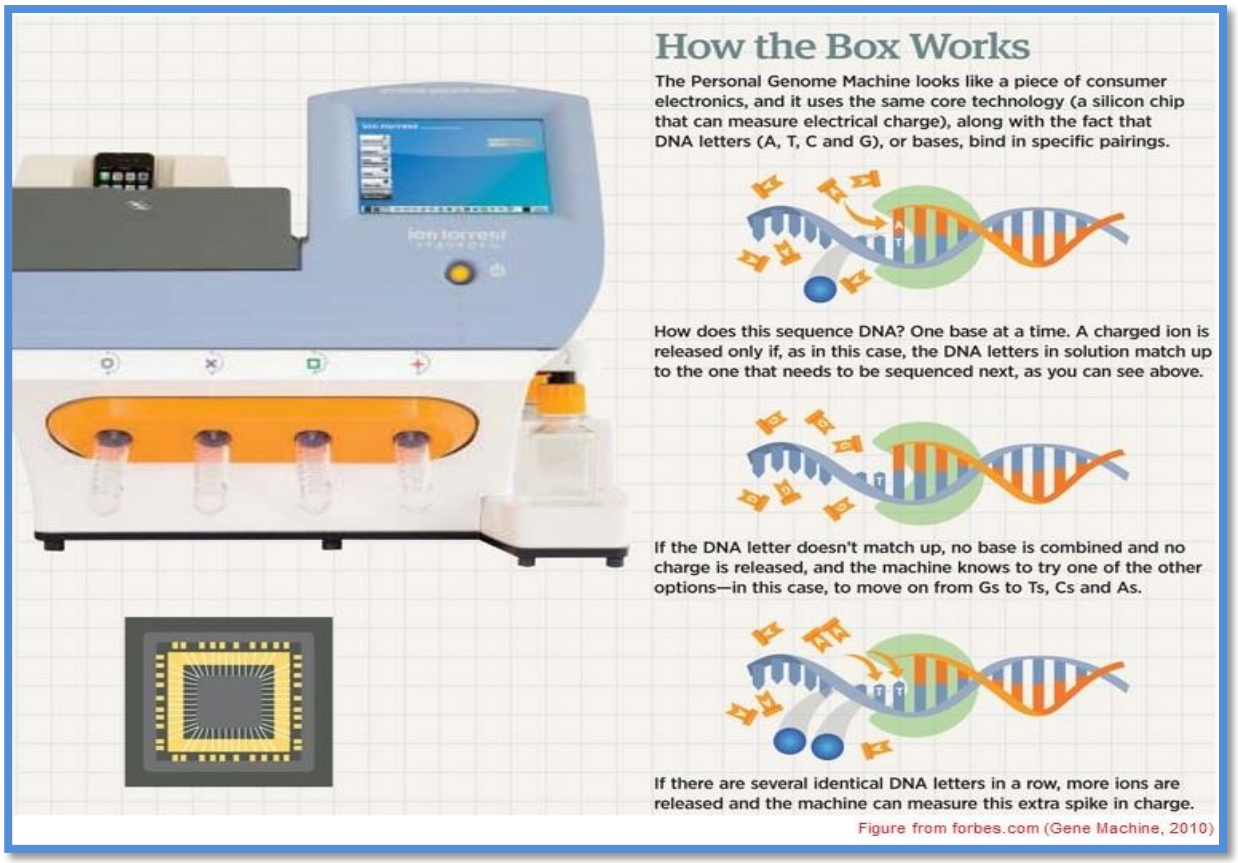


Figure 15 - Ion-Torrent PGM utilises semi-conductor sequencing chemistry where loaded DNA samples are supercharged with ionic electrical charges prior to sequencing. Additions of DNA bases then release a charged ion one at a time which causes a spike in the pH gradient characteristic of a particular base. The pH changes are detected and the relevant base is called by the instrument.

2.7 Metagenomic data analysis

Approximately 2 Gb of raw reads were recovered from each trial on the Illumina MiSeq and Life Technologies Ion PGM machines. Metagenomic data was downloaded via temporary FTP websites generated by NZGL, to a hard disk before being subjected to extensive analyses. Large files in fastq format were processed through standard quality checks and a primary analysis process before being annotated and classified according to the make-up of their communities for further comparison and functional studies.

2.7.1 Pre-processing the metagenomics raw reads

Metagenomic reads were quality checked prior to contiguous sequences assembly, comparison and annotation (Figure 16). For data processing, all raw data sequences (Illumina and Life Technologies platforms) were trimmed for sequencing adapters via 'CutAdapt' software (MIT, version 1.3) prior to QC. For preliminary quality assessment both FastQC (Babraham Bioinformatics, version 0.10.1) and SolexaQA 2 (Massey University, version 2.2) software were used to analyse the trimmed reads (Figure 16). For SolexaQA 2, sequencing reads were filtered with quality cut off values P = 0.01 and 0.05 (less than 1% and 5% error rates) before being displayed in a matrix line chart and a heat map for visual representation of quality sequences.

FastQC software employs a simple windows user interface where numerous output files (fastq, BAM, SAM) can be loaded for quality checking and the results are posted in a user-friendly HTML-based graphical reporting format. The reporting parameters contain the basic statistics parameters for analyzing the raw sequences such as the base sequence quality, sequencing quality scores, base sequence content, base GC content, sequence GC content, base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and finally Kmer content.

For FastQC analysis, a standard parameter setting (P = 0.01) was loaded by default during sequence per base quality assessment, where an acceptable quality of nucleotides was generated and displayed via HTML webpage figures, once analysis was completed (Figure 16). Next, the raw sequences (Illumina only) were filtered by a process known as 'demultiplexing' according to their unique 7 bp indexing nucleotide tags position using the FastX toolkit (Affero GPL, version 0.0.13.2) (Figure 16). There was no demultiplexing

process required for the Ion-Torrent data as the sample was run individually on a single Ion 318 conductor chip. For trimming, metagenomic data was subjected to sets of itinerary checks to remove duplicates, replicates, bad-quality bases sequence reads of more than 100 bp (trimmed back). These procedures were completed by using the following programmes, DynamicTrim and LengthSort (Massey University, version 2.2) and FastQ/A Trimmer (Affero GPL, version 0.0.13.2) (Figure 16). All metagenomic data were trimmed back to 100 bp for equal representation of read lengths across all of the different platforms and methodologies used for these comparative studies. Only high quality sequence reads (phred score of > Q30, P = 0.01) were retained for further computational analyses (Figure 16). After trimming, the raw reads were re-analyzed again using the QC software (as above) for better representation or visualisation of metagenome data.

Pre-processing workflow

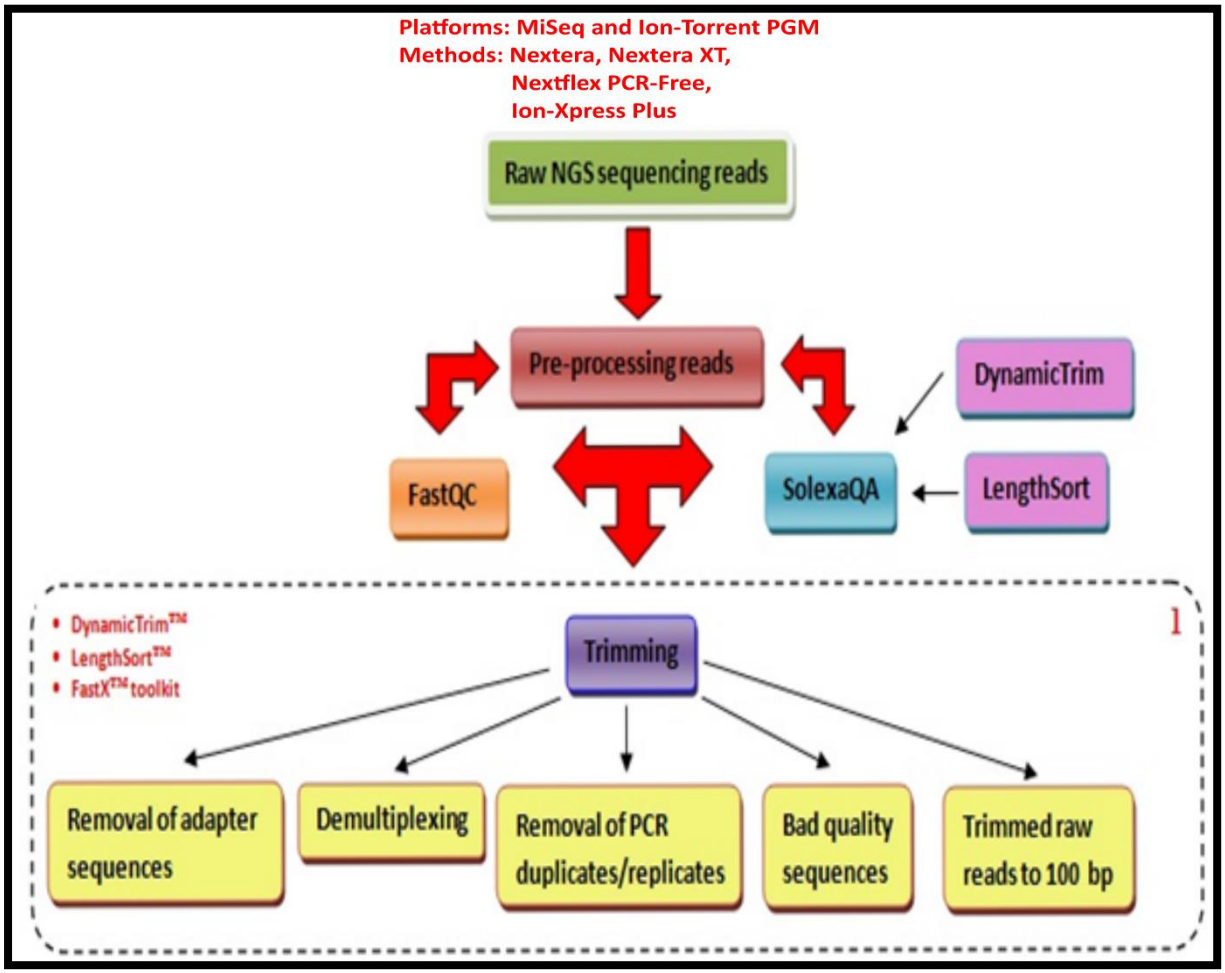


Figure 16 - Raw reads produced from different platforms and methodologies were pre-processed: the metagenomic data was quality checked, filtered, trimmed and binned before taxonomic classification and annotation.

2.7.2 Primary Data Analyses

PAUDA-Build workflow

For classification algorithms, the metagenomic reads were converted to Fasta file format using a 'Fastx toolkit' FastQ to Fasta converter (Affero GPL, version 0.0.13.2). The primary analyses consisted of a two-step process, firstly: the metagenomic sequencing reads were split into smaller query files (reduction) which speeds up the query searching and matching of sequences against public databases; secondly, the smaller reads were blasted against the NCBI-nr (non- redundant protein sequences) database using PAUDA (Protein Alignment Using DNA Aligner) (Figure 17). This software was downloaded from the website of the Bioinformatics Department, University of Tübingen (http://ab.inf.uni-tuebingen.de/data/software/pauda/download/welcome.html) to a local computer drive before being run under a high-performance built Linux based system server at Massey University, Palmerston North.

Prior to blasting, the NCBI database was indexed, before being aligned to the DNA reads. This binning process consisted of a two-stage phase: PAUDA-build and PAUDA-run (Figure 17). For PAUDA-build, the NCBI-nr database was firstly downloaded and converted to pDNA sequences (protein2pna) with an index, using a customized script modified by Dr Patrick Biggs (Huson & Xie, 2014) via Bowtie2 DNA aligner (Langmead B, Salzberg S, 2012, version 2.1.0) (Figure 17). For PAUDA-run, the sequencing reads were first translated to pDNA sequences (dna2pna) prior to filtration to avoid low complexity sequences (Huson & Xie, 2014, Wootton and Federhen, 1993) (Figure 17). Here, our sequencing reads from the MiSeq data (~3 million reads) and Ion-Torrent PGM data (~5 million reads) were binned and aligned for PAUDA metagenomic analysis. Next both builds were run together using Bowtie2 aligner software for comparison of pDNA sequences against the PNA database index. Once the alignment process was completed the pDNA reads were converted to BLASTX format using the in-built Bowtie2 aligner pnatoblastx to make an output file (Figure 17). Thereafter the results of the BLASTX alignments were generated based on the probability of the highest number of sequence similarity hits of the metagenomic reads against NCBI-nr databases. The generated output files were imported into MEGAN and saved as RMA files (read-match archive). These RMA files contained the NCBI taxonomy identification information and could be analyzed using comparative metrics in MEGAN5 software

PAUDA-Build workflow

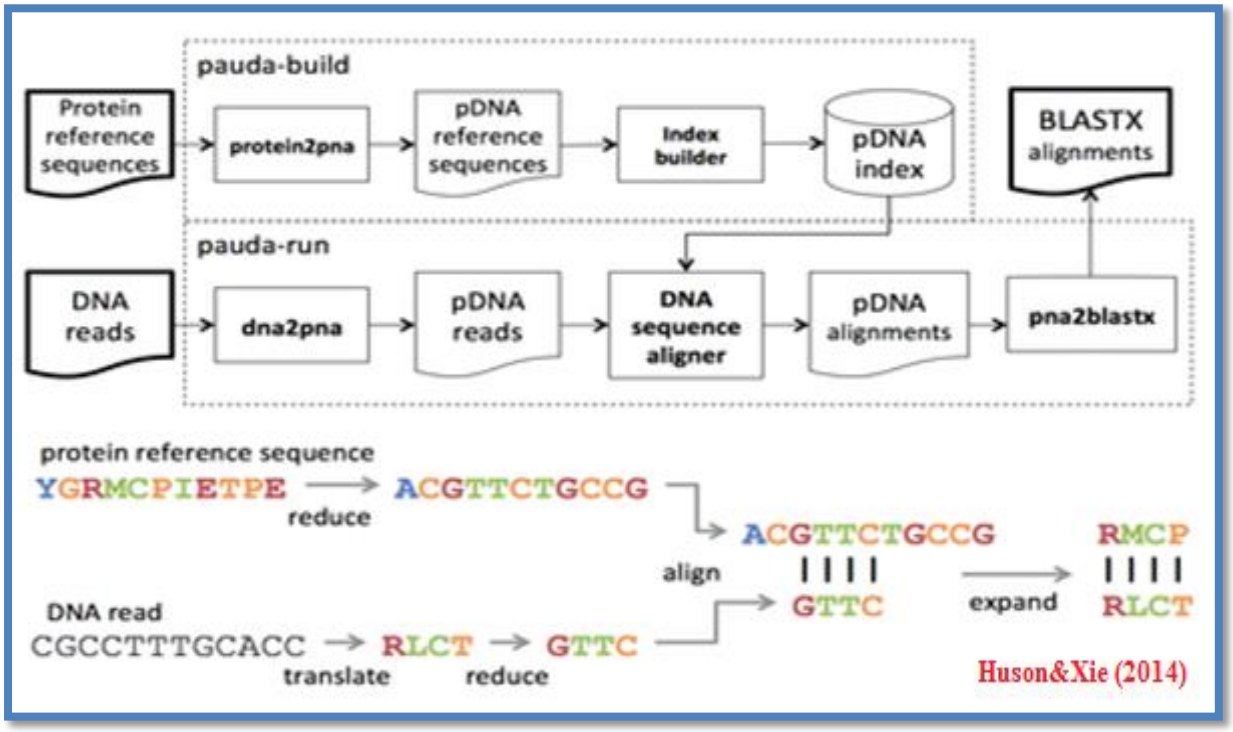


Figure 17 - PAUDA analysis, protein reference sequences from the NCBI nr database are pre-processed with index code (pDNA) for computational analysis (PAUDA-build) before alignment with DNA reads (PAUDA-run) prior to generating outputs as BLASTX alignments.

2.7.3 Comparative Outputs and Functional Analyses

The metagenomic data were compared in two ways. Firstly, we compared the data output of different NGS platforms, costs and different library preparations protocols. Here the associative costs such as the sequencing machine, running cost (cost/Mb), sequencing method, run time, read length, total reads, library insert size, DNA yield and quality were compared. Secondly we utilized MEGAN5 for community and functional analyses of the sequencing data where the pro's and con's of the sequencing methodologies will be discussed.

For data analysis from different NGS platforms (MiSeq and Ion PGM) and sample preparation methods (Nextera, Nextera-XT, NEXTFlex PCR-free and Ion-Xpress 400bp), matching reads from PAUDA analysis were used to construct the RMA files in MEGAN5 which could be analyzed later. The creation of these files was an essential step for further analyses of both taxonomic and functional identifications in MEGAN5 using SEED classification and KEGG pathways orthology (Figure 18). For SEED analysis, metagenomic

reads from the assigned taxonomy IDs (obtained using the LCA algorithm in MEGAN5) were rooted into a new classification for functional analysis (Figure 18). This SEED-based classification has different categories of "multi-labelled" functional nodes and each node is further divided into different types of functional roles e.g., carbohydrate, protein, nitrogen metabolism and many others, up to 10,000 nodes in total (Figure 18). Here approximately 10 to 20 Gb of generated Tamaki River data in RMA files were loaded for SEED classifications using minimum LCA scores of 50, filtered with the top 10% of microbial populations using a p-value of 0.01 (error rate). Meanwhile for KEGG analysis, Tamaki River metagenomics reads with matching scores of 35 and above (p-value 0.01) were selected and assigned to their respective KEGG metabolic pathways with respective KEGG orthology number. Lastly, for bacterial identification, MEGAN5 assigned the highest matching reads in PAUDA analyses against a protein database prior to the annotation of the read sequences with their putative biological roles (Figure 18).

Meta-data workflow

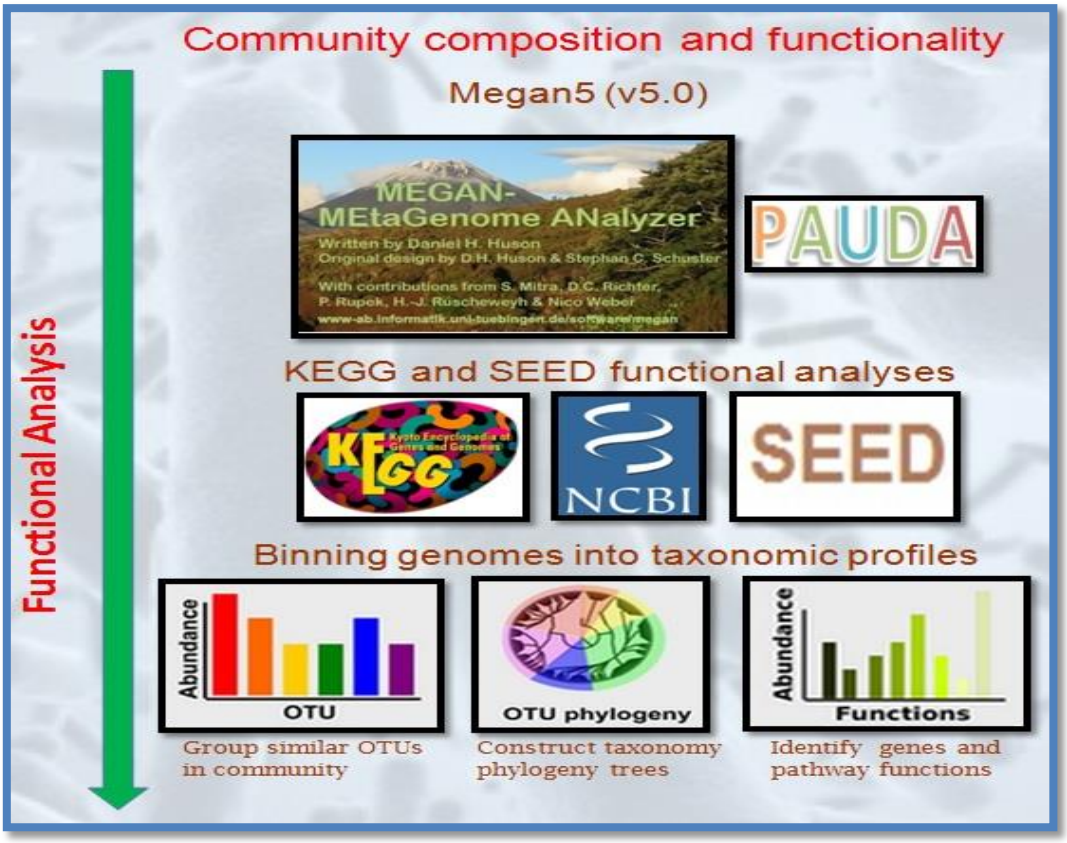


Figure 18 - Functional analyses of meta-data workflow. Blasted NCBI PAUDA data were loaded into MEGAN5 for both SEED and KEGG analyses to investigate their biological roles and also to group them together to identify different clusters of genes and functional metabolic pathways.

3 Results

3.1 Microbiological tests conducted for water quality

Table 1 shows the results of five randomly collected water samples from the Tamaki River screened against *Giardia* and *Cryptosporidium* in a colorimetric test, microscopy screening and a PCR detection test. Sample no. 3 tested positive for *E.coli* while 2 out of 5 grab samples (sample 4 and 5) were positive for *Giardia* and *Cryptosporidium*. However *Giardia* and *Cryptosporidium* cysts/oocysts were not detected by microscopy in any of the collected samples. The samples were then pooled together at the end for next generation sequencing.

Microbiological tests conducted on Tamaki River samples

Name (s)	***Detection of Pathogens (*Cryptosporidium/Giardia*)**		**Colorimetric test**	**Number of cysts/oocysts per 50 litre**	**Water Temperature (° C)**	**Water pH**
	Microscopy	**PCR**				
1 (Tamaki)	-/-	-/-	Passed	nil	6.7	8.22
2 (Tamaki)	-/-	-/-	Passed	nil	7.6	8.16
3 (Tamaki)	-/-	-/-	Failed	nil	7.1	8.12
4 (Tamaki)	-/-	+/+	Passed	nil	7.6	8.54
5 (Tamaki)	-/-	-/+	Passed	nil	7.3	7.94

Table 1 – Screening for *Cryptosporidium*, *Giardia* and *E.coli* in Tamaki River grab samples collected in November 2011. Only one sample (number 3) tested positive for coliform bacteria with the colorimetric test. Samples 4 and 5 tested positive for *Cryptosporidium* and *Giardia* with the PCR test.

Filtration of water samples

In order to determine the pore-size of filters that would yield the best results for filtering grab samples from the Tamaki River, we initially experimented with one-litre grab water samples from the duck pond at Massey University, Palmerston North. We used 1.0, 0.8, 0.45, 0.22 and 0.1 µm filters to determine which filters gave the highest yield of genomic DNA. We found that filters with pore sizes of 0.45 and 0.22 µm recovered the highest molecular weight (hmw) DNA. Figure 19 shows extracted hmw DNA from the five different pore size filters electrophoresed on 1% (w/v) agarose gel for an hour. Most filters yielded a low amount of hmw DNA except for filter sizes 0.45 and 0.22 µm.

Investigation of filter pore-size efficiency on recovering hmwt DNA

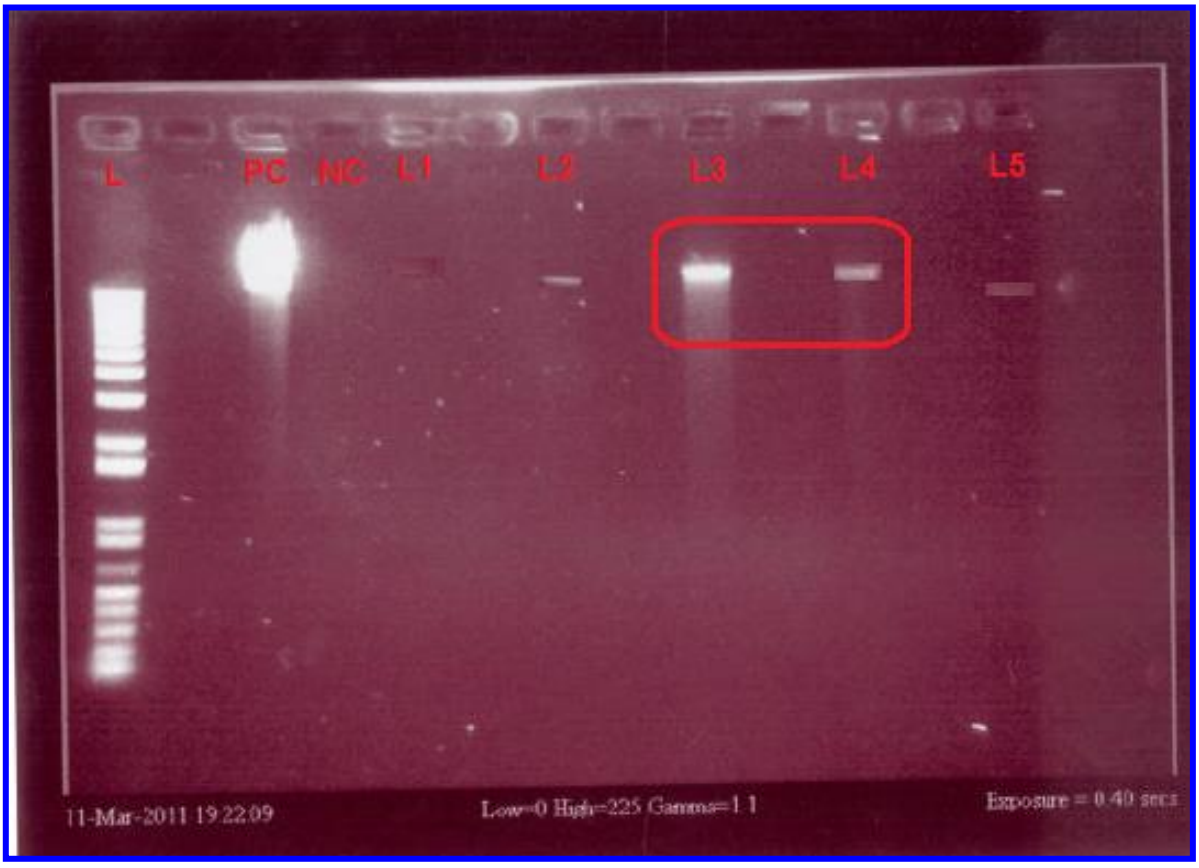


Figure 19 – Gel electrophoresis of DNA extracted from filters with different pore sizes of 1.0 µm, 0.8 µm, 0.44 µm, 0.22 µm and 0.1 µm. From left, L = 1Kb+ ladder, PC = Positive control (*E.coli*), NC = Negative control, L1 = 1.0 µm filter, L2= 0.8 µm filter, L3 = 0.45 µm filter, L4 = 0.22 µm filter and L5 = 0.1 µm filter. The red box indicates the filters we chose for our protocol.

3.2 DNA extraction

We tested the efficiency of the Epicentre metagenomics filter DNA extraction protocol, by spiking *E.coli*, ~5 µg, into one litre of clean Milli-Q water (ultra-purified, free of pathogens) as a positive control sample against a one litre grab duck pond water collected from Massey University lake (Duck Pond) as a 'real metagenomics sample'. Next we filtered and extracted DNA from these samples. Water samples were filtered through 0.45 and 0.22 µm pore size filters and DNA extracted using the Epicentre suggested extraction protocol (Figure 20). We successfully recovered hmwt DNA from all filtrates using the above filters. The concentration of DNA recovered from the *E.coli* spiked Milli-Q water was much higher in comparison to that recovered from the non-spiked duck pond water and this was expected as a positive control. However the amount of DNA was significant lower than was expected as we only managed to recovered ~2.5 µg out of 5.0 µg (a loss of ~50%) of DNA from the filters. This suggests an inefficiency in using a single filter paper for filtration as the filter can be clogged up easily thus slowing down the procedure and also limiting the amount of DNA being extracted. Besides we also observed some slight degradation (smearing) of the extracted DNA from the *E.coli* spiked duck pond water (Figure 20).

Validation of DNA extraction technique

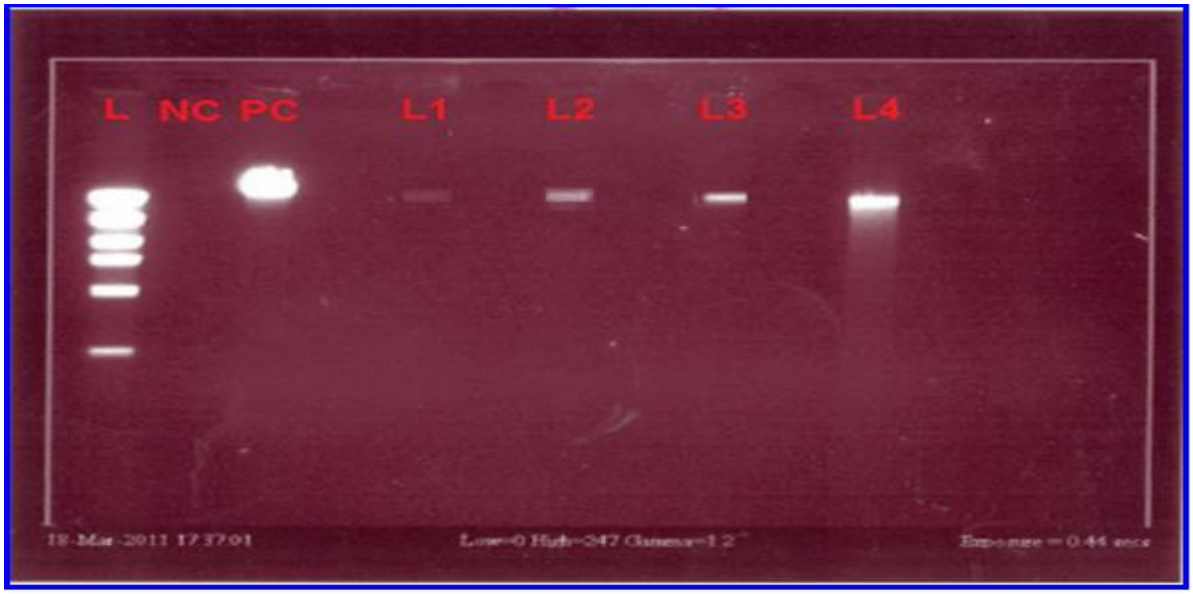


Figure 20 - Gel electrophoresis of hmwt DNA from the duck pond water (Massey University, Palmerston North) and a positive control of *E.coli* at 5 µg. From left, L = High molecular weight DNA mass ladder, NC = Negative Control, PC = Positive control (*E.coli*), L1 = 0.45 µm filter (duck pond water), L2 = 0.45 µm filter (*E.coli* 5 µg + Milli-Q water), L3 = 0.22 µm filter (duck pond water), L4 = 0.22 µm filter (*E.coli* 5 µg + Milli-Q water).

3.3 Optimisation of water filtration and DNA extraction protocols

Water samples collected from the Tamaki River were often murky, cloudy and brown in colour due to dissolved organic material and mud. High turbidity and poor clarity in a water sample have a lower success rate in DNA filtration due to clogging filters thus making it difficult to extract organic material. Here, we reviewed and optimized our filtration and extraction protocols by using multiple filters for each one litre grab sample (Tamaki River) and extracted DNA from each filter separately prior to pooling to gain maximum DNA yield from each sample. We exchanged a 0.45 µm filter for every 300ml of the one litre grab sample (up to three times for each sample) and the final eluates were filtered once again through a single 0.22 µm filter. All filters were washed using the Epicentre Metagenomics DNA Isolation kit (now Illumina Inc). The multiple filters approach yielded ~5.1µg of hmwt DNA from each one litre grab sample compared to only ~2.5µg if using only a single filter (Figure 21) (Table 2). The use of multiple filters greatly increased the speed of the filtering and also the efficiency in extracting the DNA. In the preliminary experiments with the Tamaki River water we found that the concentration of the extracted genomic DNA obtained using multiple filters was significantly higher in yield and in quality compared to that obtained using a single filter (Figure 20, duck pond water). A single filter tended to clog up and it was harder to wash the microbes off with the lysis buffer due to the thick layer of organic material stuck to the filters. The use of multiple filters also produced DNA with less degradation than the single filter. This was indicated by electrophoresis which showed that the quantity and quality of extracted DNA was greatly increased and improved when using multiple filters. The extracted DNA from both filtrates appeared as a single large hmwt DNA band of more than 40 kb (Figure 21). No band was detected in the negative control and a strong ~40kb band was detected in the positive control. Comparison of post-nebulization bioanalyzer profiles for the sonicated DNA extracted from single and multiple filers also suggested a similar conclusion. That is, the DNA peak in profiles for multiple filters consistently produced a less diffuse distribution than did the DNA from the single filters.

Gel of extracted DNA from Tamaki River

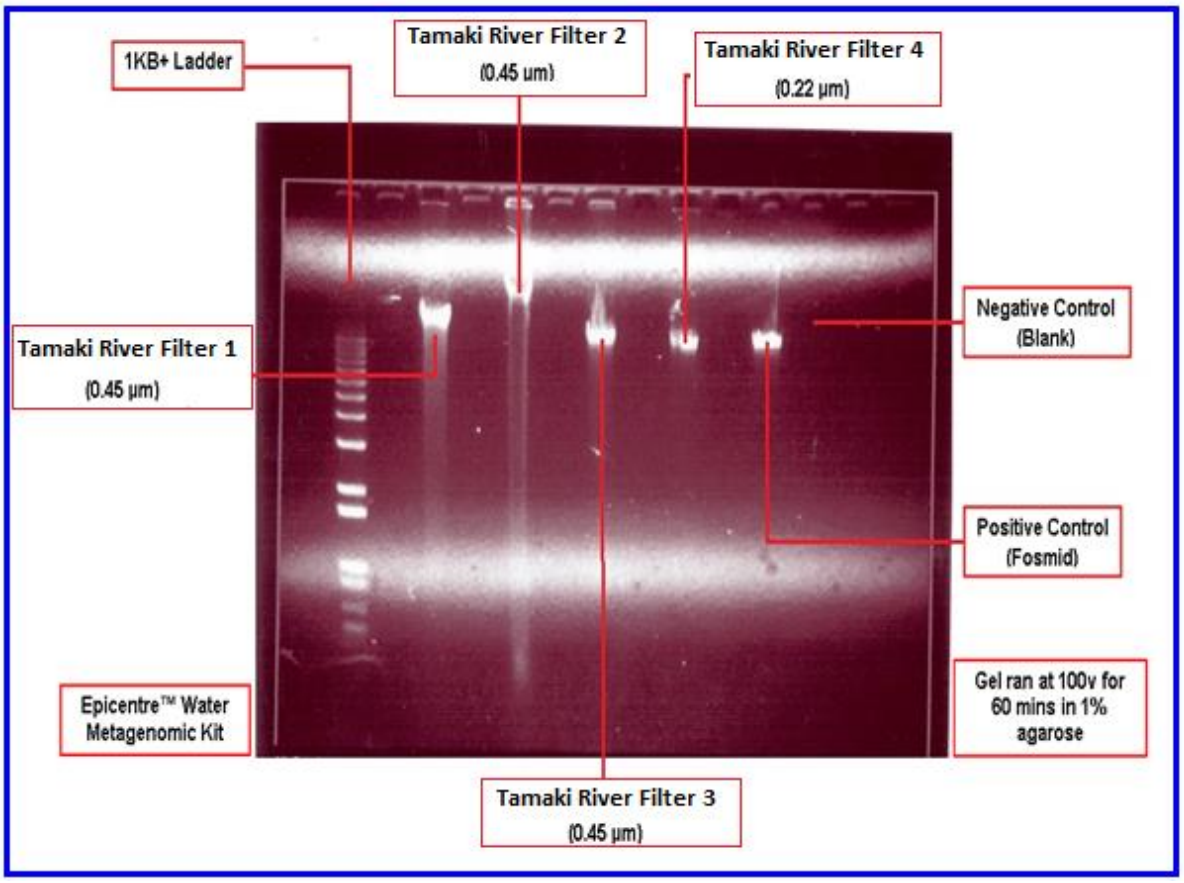


Figure 21 - DNA extracted from multiple filters (3 x 0.45 and 1 x 0.22 µm) together with positive and negative controls on a 1% (w/v) agarose gel. The band intensity for each of the filters can be compared to the previous gel (single filtration, Figure 20). The amount of DNA recovered was similar across all 4 filters. DNA from filters 1 and 2 appears to be running at a higher molecular weight compared to the DNA from filters 3 and 4 and the positive control. This result could be due to salt, or other contaminants in the final elution. All gel wells were loaded with 2 µl of purified DNA product.

To further elucidate the efficiency of the multiple filtration setup, Qubit measurements for dsDNA concentration for three-multiple 0.45 µm filters were examined (Filter 1, 2 and 3 in Figure 21). These gave readings of 1548, 1462.5 and 1471.5 ng/µl respectively with Nanodrop OD260/280 (DNA purity) measurement reading at 1.80, 1.81 and 1.91 (Table 2). By way of comparison, the concentration of DNA obtained from a single 0.22 µm filter was at 612 ng/µl with a DNA purity reading of 1.86 (Table 2). The DNA purity reading for all prepared samples were within the recommended NGS quality specification from 1.80 to 2.00. In addition Figure 22 also successfully shows the bioanalyzer profile (DNA 1000 Chip) of the fragmented library with an average fragment peak size range from 400 to 800 bp. This is an

ideal fragment for the construction of NGS library. In general, a higher quality and quantity of starting material is generally desirable for the construction of NGS libraries protocols.

Bioanalyzer profile for multiple-filtration protocol, Tamaki River

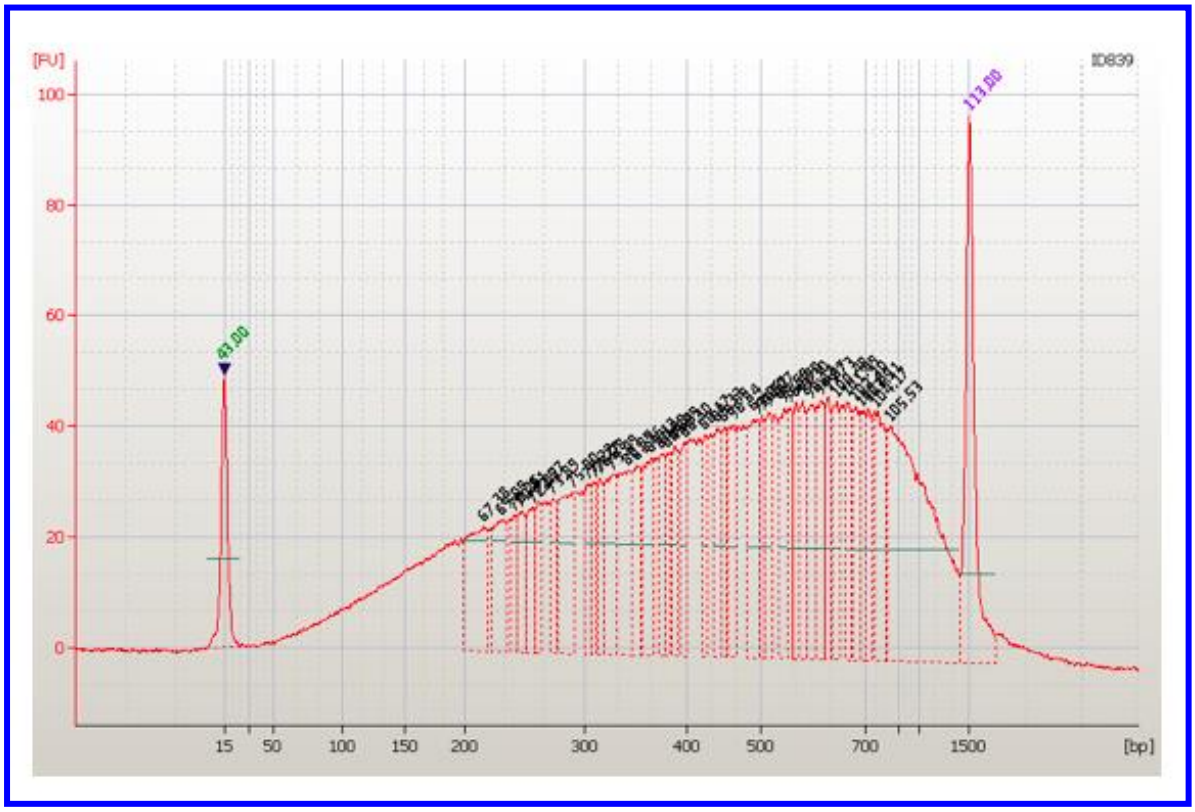


Figure 22 - Fragmented genomic DNA from a multiple filtration protocol. The sheared genomic DNA was within the recommended DNA peak size range of 400 to 800 bp which indicates that the fragmentation process had been successful and is suitable for MiSeq paired-end sequencing. FU: arbitrary fluorescent unit.

Quality and quantity of DNA obtained using single and multiple filters

Quality assessment (QA)	Single-filtration		Multiple-filtration			
Filter Size (µm)	0.45	0.22	0.45	0.45	0.45	0.22
Sample concentration per tube (ng/µl)	63.2	0.447	34.4	32.5	32.7	13.66
DNA purity (OD 260/280)	1.79	1.67	1.80	1.81	1.91	1.86
Total volume per tube (ng/µl)	40	40	45	45	45	45
Total concentration per tube (ng/µl)	2528	17.88	1548	1462.5	1471.5	612
Total concentration (ng/µl)	2545.88 (~2.5µg of starting material)		5094 (~5.1µg of starting material)			

Table 2 - Quality and quantity measurement for a single filter and multiple (0.45 and 0.22 µm) filters. We used both Qubit and Nanodrop instruments for this assessment. Most of the sample purities were within an acceptable range for the construction of a NGS library (1.8 to 2.0). The concentration of DNA in the 0.22 µm final pooled samples from multiple filters was significantly higher compared to that obtained with single 0.45 µm filter.

3.4 Metagenomic library preparations

3.4.1 Nextera and Nextera-XT DNA Library Construction

As shown in Table 2, pooled samples from multiple filtrations (Tamaki River) were chosen for the Nextera library protocol due to better quality material and are within the recommended DNA purity specification (1.80-2.00) with an average OD 260/280 of 1.84. Following digestion, end repair, ligation of adaptors and PCR enrichment, the size of the amplified (library) fragments was determined on the Bioanalyzer. The Bioanalyzer profiles from both library preparation showed we have an average fragment size of 485bp for the Nextera protocol (Figure 23) and 510bp for the Nextera-XT protocol (Figure 24). In addition, the bioanalyzer profiles also showed that it had approximately 10.3ng/µl (Nextera) and 6.87ng/µl (Nextera-XT) of DNA for each of the prepared libraries. Further evaluation and quantification of both libraries by Qubit assays for DNA (high sensitivity), RNA and Protein content revealed no significant contamination from RNA (<20ng/mL) or protein source (<1.0

ng/mL; Table 3). Meanwhile the Qubit dsDNA high sensitivity assay showed a DNA concentrations of 8.3ng/µl for Nextera and 5.7ng/µl for Nextera-XT library (Table 3) which were similar to the concentrations estimated by the Bioanalyzer. Therefore both NGS libraries met the minimum concentration requirements (2nM of constructed libraries) for the Illumina MiSeq sequencing protocol prior to cluster generation and sequencing.

Size distribution for PCR-enriched Nextera library

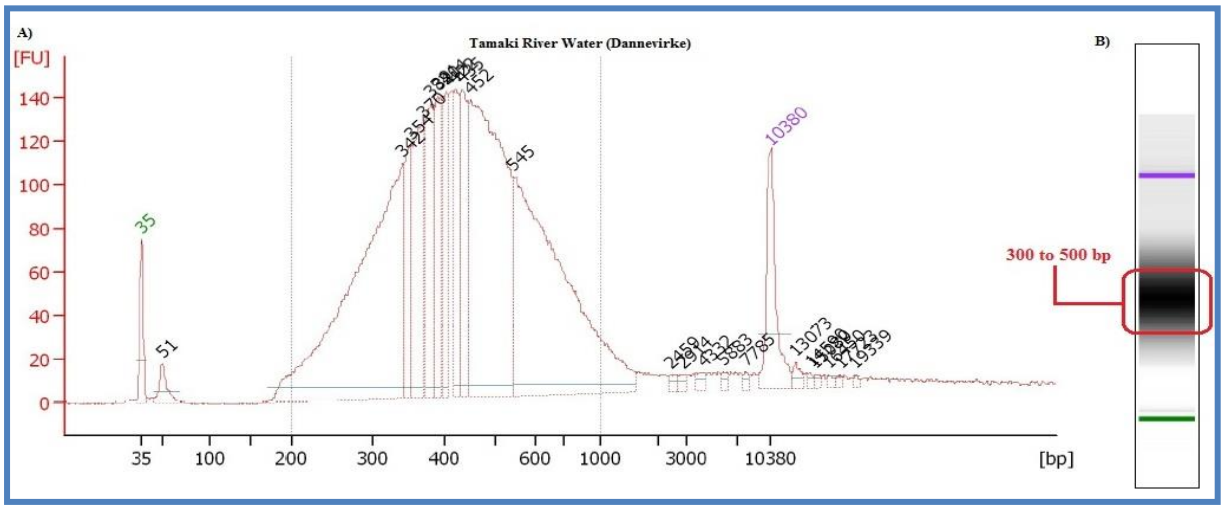


Figure 23 A) Bioanalyzer profile for Nextera libraries indicating fragment size range of 200 to 800 bp following PCR enrichment and B) gel view showing most of the fragments were between 300 to 500 bp. FU: arbitrary fluorescent unit.

Size distribution for PCR-enriched Nextera-XT NGS library

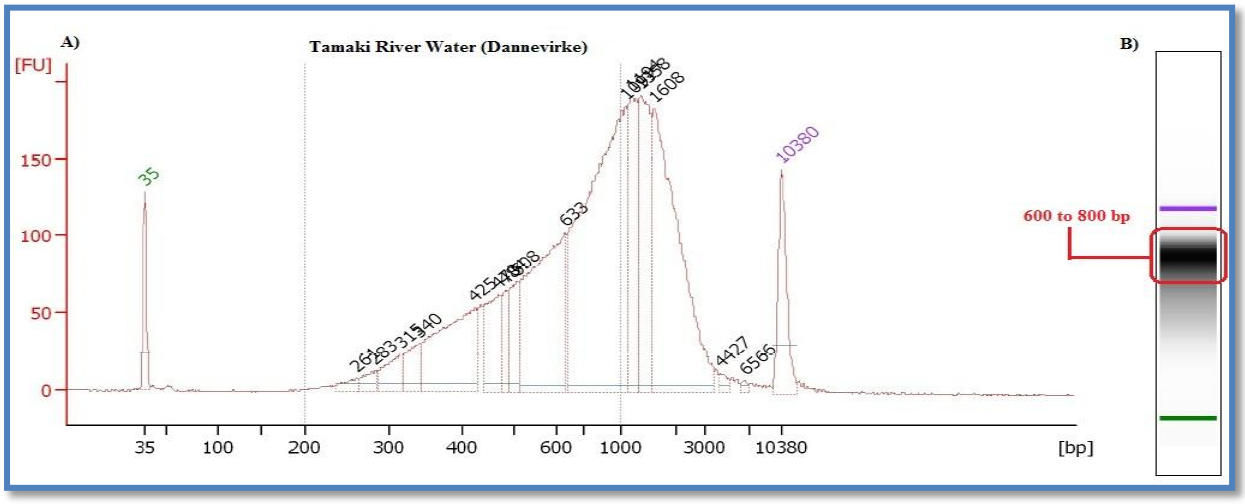


Figure 24 A) Bioanalyzer profile showing that the Nextera-XT library fragments were larger than those obtained with the Nextera procedure and B) gel visualisation indicating most amplified products between 600 to 800 bp. FU: arbitrary fluorescent unit.

Quantification of Nextera and Nextera-XT libraries

Assay	Concentration in the Qubit	uL used	Dilution	Sample Concentration
Quant-iT Protein	Out Of Range <1.0 ng/mL	2		-
Quant-iT RNA	Out Of Range <20 ng/mL	2	200	-
Quant-iT dsDNA	255 ng/mL	2	200	51 ng/mL

Table 3 – Qubit quantification readings obtained from Qubit fluorometer for protein, RNA and DNA assays. Both Nextera and Nextera-XT libraries showed an acceptable level of protein and RNA (less than 1 ng/µl) with total DNA concentration of 51 ng/µl.

3.4.2 NEXTFlex PCR-free DNA Library Construction

Prior to PCR-free library preparation, the Tamaki River DNA was again checked on a 1% (w/v) agarose gel to ensure the DNA was of the same integrity as used for the other Illumina library preparations (Figure 25). Next, 23.53 µl (1.2 µg) of this sample was mixed with 26.47 µl of sterile water (to a total volume of 50 µl), for fragmentation by nebulisation. The Bioanalyzer profile after nebulisation shown in Figure 26 indicated that most of the library products were in the size range of 350 to 700 bp, which represents a size suitable for sequencing with this protocol.

Genomic DNA used for NEXTFlex PCR Free library construction

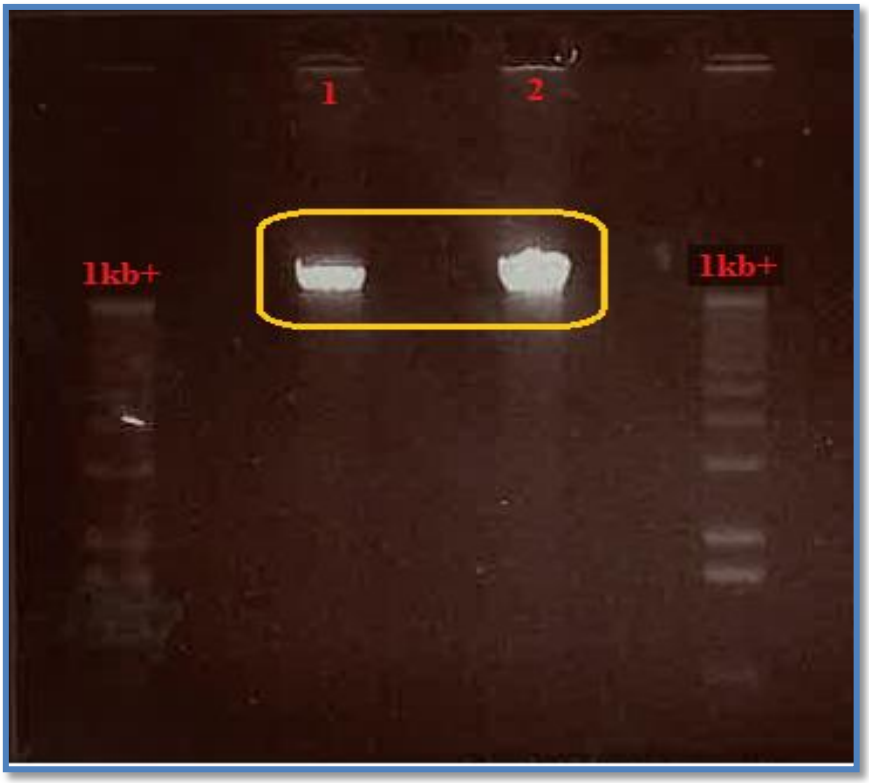


Figure 25 – A total of 5 µl of genomic DNA from Tamaki River was loaded into lanes 1 and 2. We observed the presence of hmw DNA (yellow square) in both lanes. Both bands are strong with minimal degradation.

Size distribution for NEXTFlex PCR-free Library

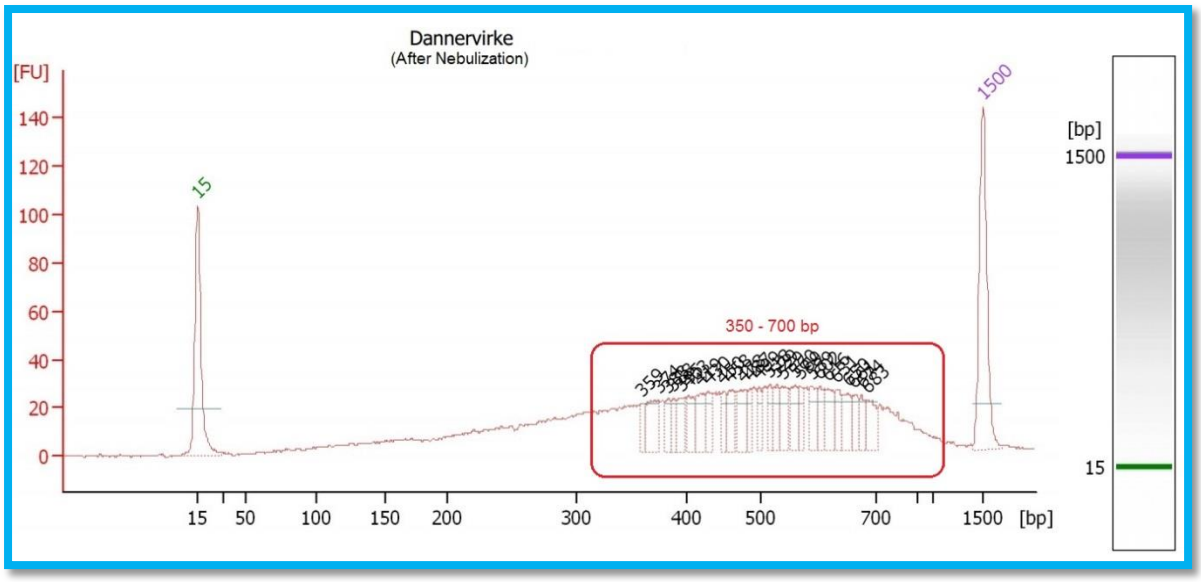


Figure 26 – Bioanalyzer DNA 1000 profile of the PCR-free protocol showing size distribution of the library fragments. These ranged between 350 and 700bp. After nebulization the concentration of the total amount of DNA dropped from 1.2 µg to 0.93 µg. FU: arbitrary fluorescent unit.

The overall DNA yield from the PCR-free based protocol was successful, but the library was of a lower concentration to that obtained from both Nextera protocols. This was to be expected as there was no PCR enrichment step in the library preparation protocol to eliminate any PCR biases. The bioanalyzer profile indicated an average DNA library fragment size of 581bp for a concentration of 1.08ng library/µl. Further assessment with the Qubit fluorometer showed a dsDNA concentration of approximately 2.26ng/µl with minimal RNA and proteins contamination (Table 4). For MiSeq sequencing, we diluted the PCR-free library from a concentration of 2nM to 10.5pM before loading it onto an Illumina MiSeq instrument.

Quantification result of PCR-free library using Qubit

Assay	Concentration in the Qubit	uL used	Dilution	Sample Concentration
Quant-iT Protein	Out Of Range <1.0 ng/mL	2		-
Quant-iT RNA	Out Of Range <20 ng/mL	2	200	-
Quant-iT dsDNA	11.3 ng/mL	2	200	2.26 ng/mL

Table 4 – Quantification of protein, RNA and dsDNA levels made with a Qubit fluorometer. The Tamaki River sample had less than 1% RNA and protein contamination. The average library fragment size was at 581bp.

3.4.3 Ion-Torrent PGM Library Preparation

Library preparation for the Ion-Torrent platform was carried out by NZGL (University of Auckland). Approximately 1µg of hmw DNA from the Tamaki River with an OD ratio (260/280) of 1.84, was used as starting material. Using the Ion Xpress 200 bp sample preparation kit, the gDNA was enzymatically sheared to < 500 bp and end-repaired. The bioanalyzer reading following end-repair showed a concentration of 453.32 ng/µl (Figure 27A). Following adaptor ligation and size selection (using SPRI beads) the prepared sample was successfully enriched by PCR, with the majority of dsDNA fragments being ~ 200 to 300 bp in length (Figure 27B).

Ion-Torrent size distribution before and after NGS library construction

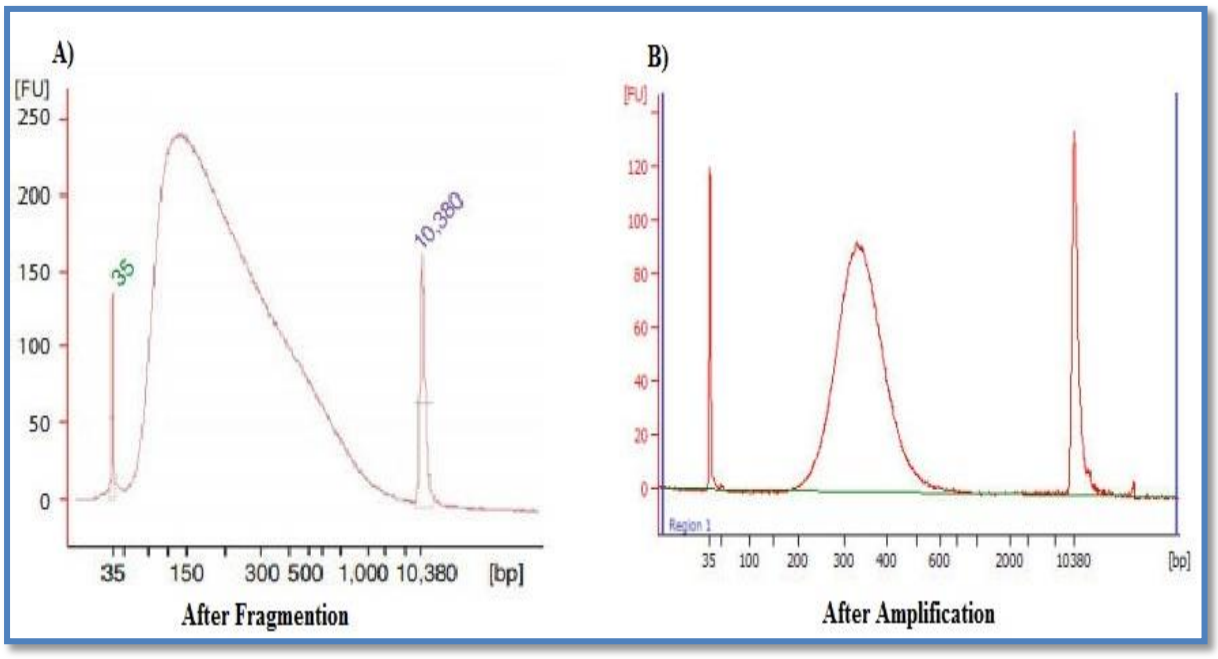


Figure 27 – A) Size distribution for 1µg of gDNA after fragmentation for 20 minutes at 25°C and 10 minutes at 70°C. After fragmentation, the majority of the DNA fragments were < 800 bp. B) Size distribution for enriched NGS library after size selection and emulsion-PCR amplification. The majority of the final library fragments were between 200 – 400 bp in size. FU: arbitrary fluorescent unit.

3.5 Next Generation Sequencing

3.5.1 Illumina Sequencing

3.5.1.1 MiSeq Sequencing System

For MiSeq Sequencing, the amount of DNA loaded into the instrument is dependent on the average NGS library fragment-size along with its final library concentration and instrument loading molarity. The loading molarity here refers to the amount of DNA required to generate an optimal cluster density for sequencing. The amount of Tamaki River DNA used for MiSeq sequencing varied depending on the result we obtained from different protocols used: Nextera (100 ng), Nextera-XT (1 ng) and PCR free protocol (2.5 µg).

For the Nextera protocol it gained an average library fragment size of 357 bp along with final library concentration at 2.74 nM and a MiSeq loading molarity of 9.5 pM (Table 5). The library was generated with an average cluster density of approximately 961 k/mm^2. With 2 x 150bp paired-end sequencing of the Nextera library, the run returned 2.19 Gb of data (13,406,580 reads). We gained a total yield of 1.03 Gb and 1.16 Gb of sequencing data from read 1 and 2 respectively with error rates of 0.51% and 0.69%. The report also showed we had > 88.4% (Read 1) and > 81% (Read 2) of paired-end reads above the phred Q_{30} quality score. For signal intensity across the flowcell, we had a total percentage of 81.9% and 82.9% of nucleotide base calling accuracy for both read 1 and 2.

Meanwhile for Nextera-XT, it had an average library fragment size of 563 bp and a final diluted library concentration of 2.85 nM (Table 6). The total cluster density for sequencing was at 1121 k/mm^2 (a slight increase over that obtained with the Nextera protocol) with 85.2% (Read 1) and 76.4% (Read 2) being above phred Q_{30} scores. The 2x150 bp paired-end sequencing run using the Nextera-XT library yielded 2.61 Gb data (16,713,891 reads) with 83% of the data above the phred Q_{30} score (Table 6). This gave a final yield of 1.34 Gb (read 1) and 1.32 Gb (read 2) of data with error rates of less than 0.41% (read 1) and 0.57% (read 2). In respect of the MiSeq focus quality score, the local SAV files showed good signal intensity representation across all nucleotide bases where 80.3% (Read 1) and 78.3% (Read 2) were above the pass filter rate.

Overall both the Nextera and Nextera-XT sequencing data were successful and comparable to that routinely obtained using the gold standard TruSeq DNA preparation protocol (data not

shown), even though a much smaller amount of good quality starting material (gDNA) was required for the NGS library construction.

Run Summary for Nextera on the MiSeq platform

	Average Library Size Fragment (Bp)	**Final Library Molarity (nM)**	**Cluster density (k/mm^2)**	**Total yield (Gb)**	**Total no of reads**	**Error Rates (%)**	**Q30 score (%)**	**% Optics intensity Cycle 20**	**Cluster PF (%)**
Read 1	357	2.74	961 +/- 11	1.03	6,512,311	0.51	88.4	81.9	88.4 +/- 0.6
Read 2	357	2.74	961 +/- 11	1.16	6,894,269	0.69	81.4	83,9	88.4 +/- 0.6
Total	357	2.74	961 +/- 11	2.19	13,406,580	0.60	85.1	82.9	88.4 +/- 0.6

Table 5 – The Sequencing Analysis Viewer (SAV) summary report indicated that we had a total data output of 2.19 Gb with less than 0.6% error rate and 99.4% base-calling accuracy. We obtained an optimal cluster density of 961 k/mm^2 with an average passing filter Q_{30} score of 82.9% for the Tamaki river water sample.

Run Summary for Nextera-XT on the MiSeq platform

	Average Library Size Fragment (Bp)	**Final Library Molarity (nM)**	**Cluster density (k/mm^2)**	**Total yield (Gb)**	**Total no of reads**	**Error Rates (%)**	**Q30 score (%)**	**% optics intensity Cycle 20**	**Cluster PF (%)**
Read 1	563	2.85	1121 +/- 10	1.32	8,512,311	0.41	85.2	80.3	92.2 +/- 0.2
Read 2	563	2.85	1121 +/- 10	1.29	8,201,580	0.57	80.3	76.4	92.2 +/- 0.2

Total	563	2.85	1121 +/- 10	2.61	16,713,891	0.49	85.1	82.9	92.2 +/- 0.2

Table 6 – A total of 2.61 Gb was generated for this run with error rates less than 0.5% and 99.5% accuracy for nucleotide base calling. We obtained a high cluster density of 1121 k/mm^2 for which 85% data was categorised as 'good quality'.

Next, a single 2 x 250 bp paired-end sequencing run of the library generated from the PCR-free protocol with an average fragment size of 901 bp and library molarity of 2.65 nM, returned 2.34 Gb (12,813,091) of raw reads data with 91% above the phred Q_{30} quality score; this was the highest accuracy of sequencing data obtained (Table 7). Next, we also gained a total cluster density of 1003 k/mm^2 with phred Q_{30} scores of 92.3% and 94.5% for both reads. For image quality we had a total signal intensity percentage (cycle 20) of 93.6% (read 1) and 95.6% (read 2).

Run Summary for NEXTFlex PCR free on the MiSeq platform.

	Average Library Size Fragment (Bp)	**Final Library Molarity (nM)**	**Cluster density (k/mm^2)**	**Total yield (Gb)**	**Total no of reads**	**Error Rates (%)**	**Q30 score (%)**	**% optics intensity Cycle 20**	**Cluster PF (%)**
Read 1	901	2.65	1003+/- 11	1.19	6,406,540	0.38	92.3	93.6	91.9 +/- 0.6
Read 2	901	2.65	1003+/- 11	1.15	6,406,551	0.22	94.5	95.6	91.9 +/- 0.6
Total	901	2.65	1003+/- 11	2.34	12,813,091	0.60	93.4	94.6	91.9 +/- 0.6

Table 7 – Run summary indicating that there was a total of 2.34 Gb of data generated from the 2x250 bp paired-end sequencing run. The table indicates a 0.6% total error rate with the final library loading molarity of 2nM . We observed a total cluster density of 1003k/mm^2 with 93.4% passing the quality filter.

3.5.2 Ion-Torrent PGM Sequencing

For comparative sequencing we chosen the Ion 318 chip (version 1) which is capable of producing approximately 1.0 Gb of sequencing data and 5.5 million reads per run, in less than 6 hours of run time. Proprietary software inherent to the Ion-Torrent platform was used to plan, coordinate, monitor and analyse the sequencing run and the final data output was exported as raw data in a fastq file, as well as data quality files, from NZGL (University of Auckland). A summary statistic generated from the run showed we had a total of 1.1 Gb of raw sequencing data and 5,512,331 million reads with average fragment read length of 147+/- 5 bp (Table 8). For read quality, we had a total 71.7 +/- 0.8% of raw reads above AQ20 (error rate of 1% or less) and mean quality score of 34.7. The majority of the 5.5 million reads passed the QC filter and contained more than 95% of adapter sequences which indicated the metagenomic sample was successfully sequenced. The Ion Sphere Particle (ISP) (beads that held the DNA) metric showed that we had an average of 82% of beads in each well of the Ion 318 chip and a high proportion of the wells were 'positive' for a pH gradient change. For quality assessment the first 300,000 reads were mapped against reference genomes and this was done internally by an NZGL bioinformatician. A mean percentage of 91 +/- 5.7% reads were mapped successfully to reference genomes with an average read mean length of 147.7 bp.

Ion-Torrent PGM summary sequencing statistics report

Sample	Raw data	# reads	% well with ISP	% with adaptors	# reads used	% reads wrapping	Mean length	Mean quality	% AQ20 quality
Tamaki River	2.01	5,428,136	82	95	300,000	95	147.7	34.7	90.9%

Table 8 – Summary statistics indicating the amount of raw data output, number of raw reads along with the percentage of wells with ISP beads. The table also shows that most reads had a length of 147.7 bp and phred quality mean score of 34.7. 90.9%. of the reads had an AQ20 read length score. These scores are similar to Phred-like scores. Here, AQ20 quality refers to a phred-like score of 20 or better, where there is one error rate per 100 bp.

3.5.3 Summary for different NGS platforms and sample preparation protocols.

For NGS libraries generated using the Illumina protocols and MiSeq instrument for the Tamaki River, the Nextera protocol generated a total of 1.93 Gb, the Nextera-XT protocol 1.88 Gb and the NEXTFlex PCR-free protocol 2.01 Gb of raw sequencing data (Table 9). Meanwhile the NGS library from Ion-Torrent PGM generated a total of 1.03 Gb of raw sequencing data (Table 9). The sequencing data produced from both different instruments and protocols were comparable in term of quality and quantity. However, given the Ion-Torrent data is only single-end read data, the raw data output is less than the data output of the Illumina sequencing.

Raw data output: Platforms and protocols

Platform	Instrument	Type of sequencing chip	Sample	Library preparation method	Raw output data (GB)
Illumina	MiSeq V2	Flowcell	Tamaki River	Nextera	1.93
				Nextera-XT	1.88
				TruSeq DNA	2.19
				NEXTFlex PCR Free	2.01
Life Technologies	Ion-Torrent PGM	Chip*	Tamaki River	Ion-Express 400 bp Kit	1.03

Table 9 – Summary of NGS raw data output from different instruments and library preparations. All NGS libraries were normalised to 2nM concentration before being loaded for sequencing.

3.6 Additional QC checks

Prior to data analysis, a quality assessment was made on data generated from the MiSeq and Ion-Torrent PGM instruments. We performed the preliminary analysis using both FastQC and SolexaQA software.

3.6.1 FastQC analysis

Data from the Illumina Nextera protocol

According to the FastQC per base sequencing quality report, raw sequences from read 1 were better quality than those of read 2. This is indicated by the greater proportion of read 2 sequences with low quality scores (Figure 28). This analysis suggested that for most of the data, both reads had less than 0.01% sequencing error rate for the first 110 base pairs. After this position in the sequence there was a significant drop in sequence quality that was most notable in read 2 (Figure 28). The analysis also showed evidence of sequence duplication at a level of 5.67% and 5.48% for reads 1 and 2 (Table 10). Sequences for read 1 and 2 failed the QC check for kmer content, indicating overrepresented DNA sequences (pentamers) present in the data. An overrepresented nucleotide repeat pattern is likely to indicate a sequencing problem. Here we observed a large spike of pentameric repetitive sequences including AAAAA, TCTCT, ATCTC, TTATA, ATACA and CTTAT at the beginning of the sequences just after ~40 bp in the raw reads (Figure 28). These pentamers are likely due to the formation of adapter dimers and activity of the 'transposase' enzyme sequences used in the Nextera shearing protocol. Such repetitive sequences are commonly reported for protocols involving transposon-based enzymatic shearing (Nextera and Nextera-XT) and need to be trimmed from the reads prior to further downstream analysis. Overall the data from the Nextera library preparations passed the above quality checks.

Data from the Illumina Nextera-XT library protocol

In addition the paired end (2 x 150bp) sequencing run for the Nextera-XT library (generated from 1 ng of gDNA) returned approximate about 3,3 million raw sequencing reads with an average sequence length of 151 bp long and a GC content of 54% for both read 1 and 2 (Table 10). Similar to the data produced with the Nextera protocol, we had a better base-calling quality result for read 1 compared to read 2 (Figure 29). With read 2 there was a gradual drop in sequence quality after a read length of 120 bp, with an average phred score of less than 20 (Figure 29). This data needed to be trimmed to acceptable quality for further

computational analysis. Both paired-ends reads had a normal GC distribution and normal sequence read lengths with some overrepresented sequences as observed with the Nextera protocol. The FastQC report showed we had a total of 6.11% and 5.7% duplicate sequences for both read 1 and 2. Similar to the data produced by the Nextera protocol there was unequal enrichment of short repetitive sequences over the read length for both paired-end reads (Figure 29). Presumably this was also due to the transposase activity during enzymatic shearing.

Sequence per base quality and kmer content in sequences generated from the Nextera protocol on a MiSeq instrument

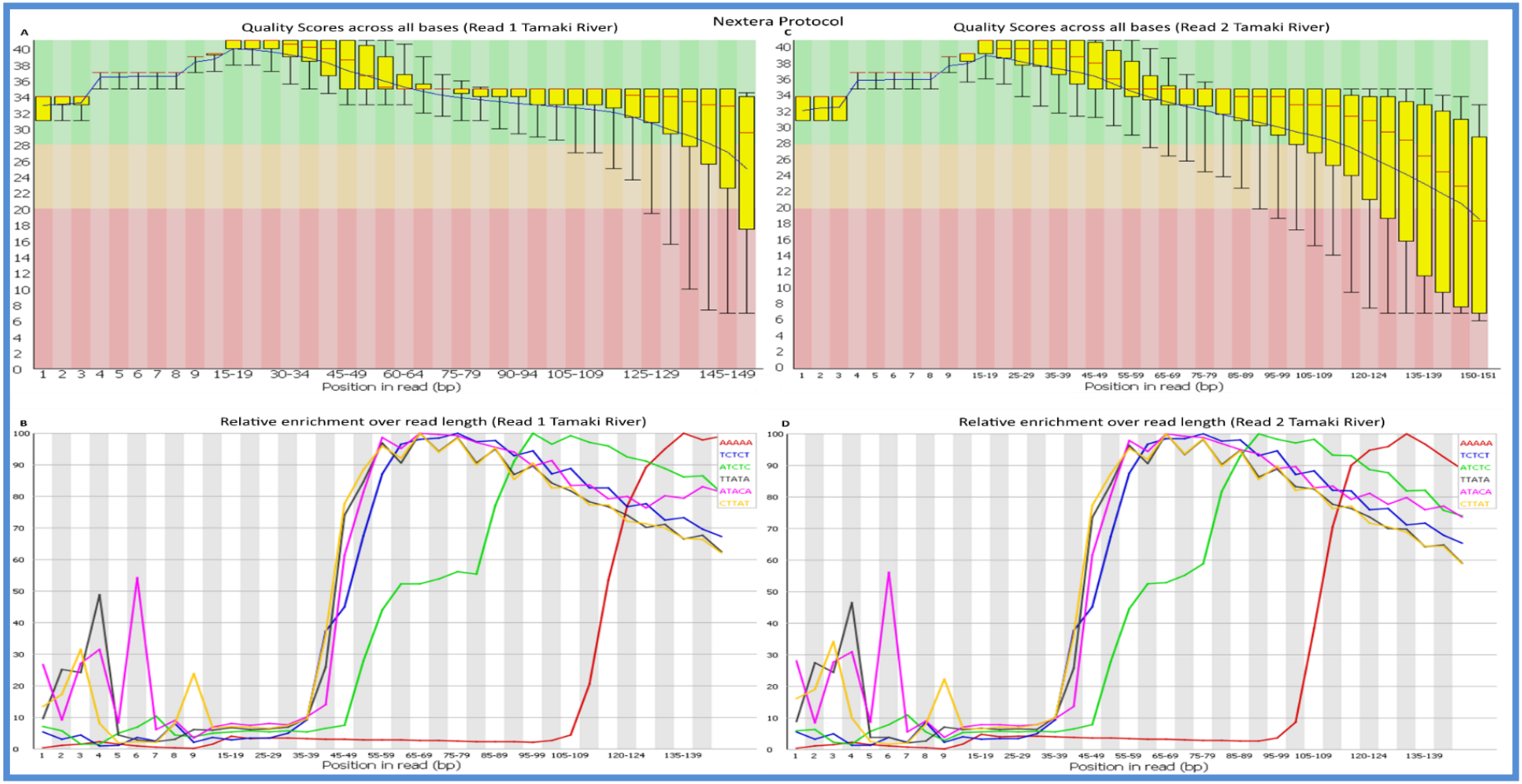


Figure 28 – FastQC analysis on sequence quality and Kmer content of read 1 and 2 data obtained with the Nextera protocol. Per base sequence quality scores for read 2 (C) were lower than for read 1 after position 120bp. For both read 1 (B) and read 2 (D) Kmer content was high at the beginning of the raw reads and also high after position 40bp .

Sequence per base quality and Kmer content for data generated from the Nextera-XT protocol on a MiSeq instrument

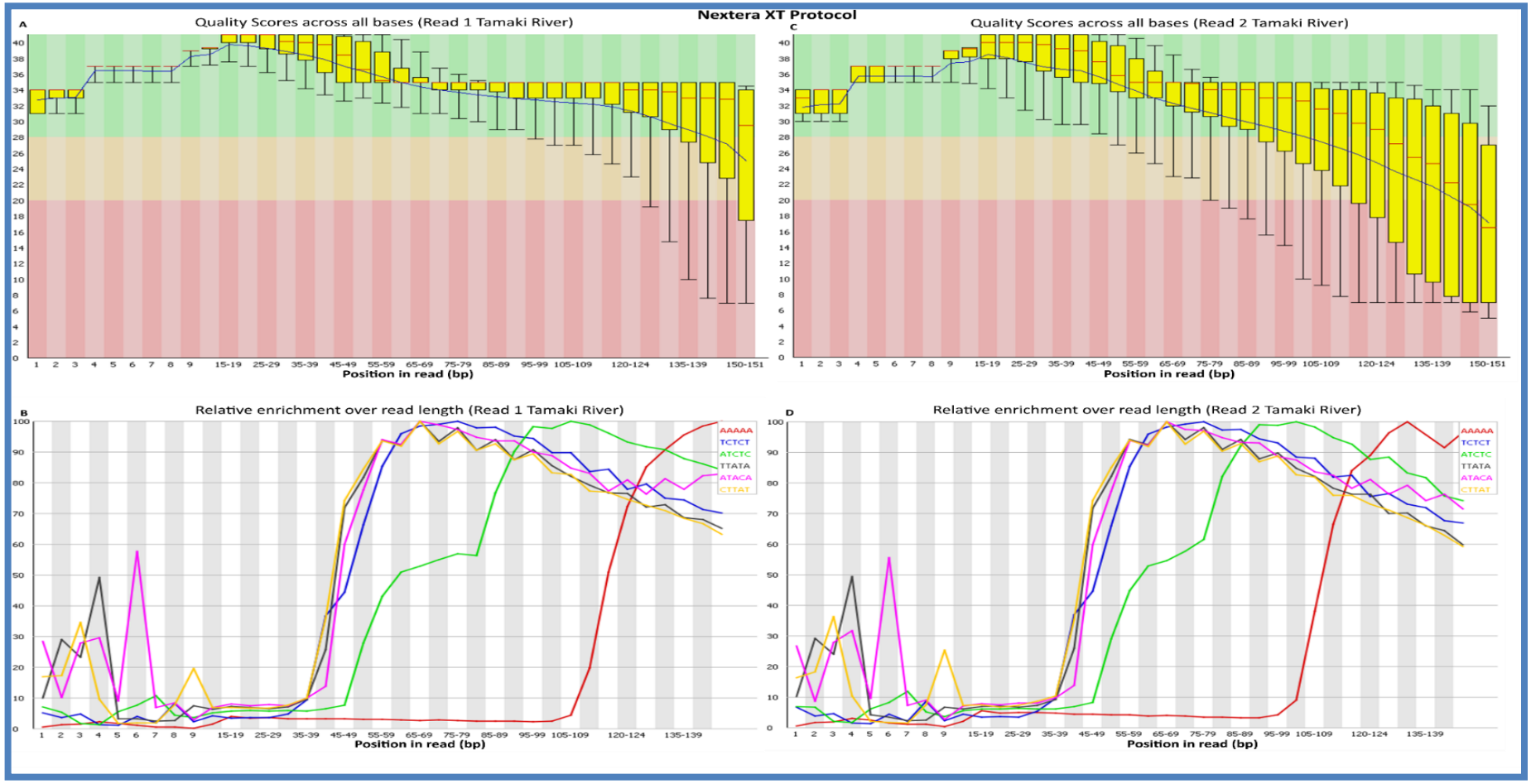


Figure 29 – Quality report for data produced using the Nextera-XT protocol. A) Quality per sequence for read 1 showed high quality with less than a 0.1% error rate. C) Read 2 showed lower quality scores compared to read 1 but still passed the quality metrics score (less than 0.1% error rate). The high kmer content likely due to Nextera-XT transposase enzyme was evident in both read 1(B) and read 2 (D).

Data from the NEXTFlex PCR-free library protocol

The PCR-free library run on the MiSeq instrument (2 x 250bp PE run) generated a total of 3,880,968 million reads with average read length of 251 bp and GC content of 53% and 54% for read 1 and 2 respectively (Table 10). We observed a high quality of data from read 1 up to 250 bp with average phred quality scores of 28 and above (Figure 30). The quality per sequence graph for read 2 showed we had quality reads (phred score > 28) up to a read length of 200 bp (Figure 30). The quality phred score dropped to 14 at position of 231 bp with an error rate of more than 10% (Figure 30). The FastQC report also showed that the data had a high GC count per read over all sequences for read 1 and 2. The report suggested we had a sequence duplication level of 11.17% and 10.58% and overrepresented sequences for both paired-end reads. The overrepresented sequences were primarily due to index-adapter contamination with 97% being over 49 bp long. This was reflected in the Kmer content chart where both read 1 and 2 showed unequal distribution of pentamerics repetitive sequences (AAAAA, TTTTT, AAAAT, TGCCG and GAAAA) over a read length of 50-80 bp. These pentamers were primer adapter sequences that were likely to have been carried over during the ligation step in the library construction. Their presence could also be due to primer dimers that were not properly isolated during the SPRI clean-up step. Generally the quality of metadata reads generated from the PCR-free protocol was of good quality and only required minor trimming.

Sequence quality and Kmer content for data generated from the NEXTFlex PCR-free protocol on a MiSeq instrument

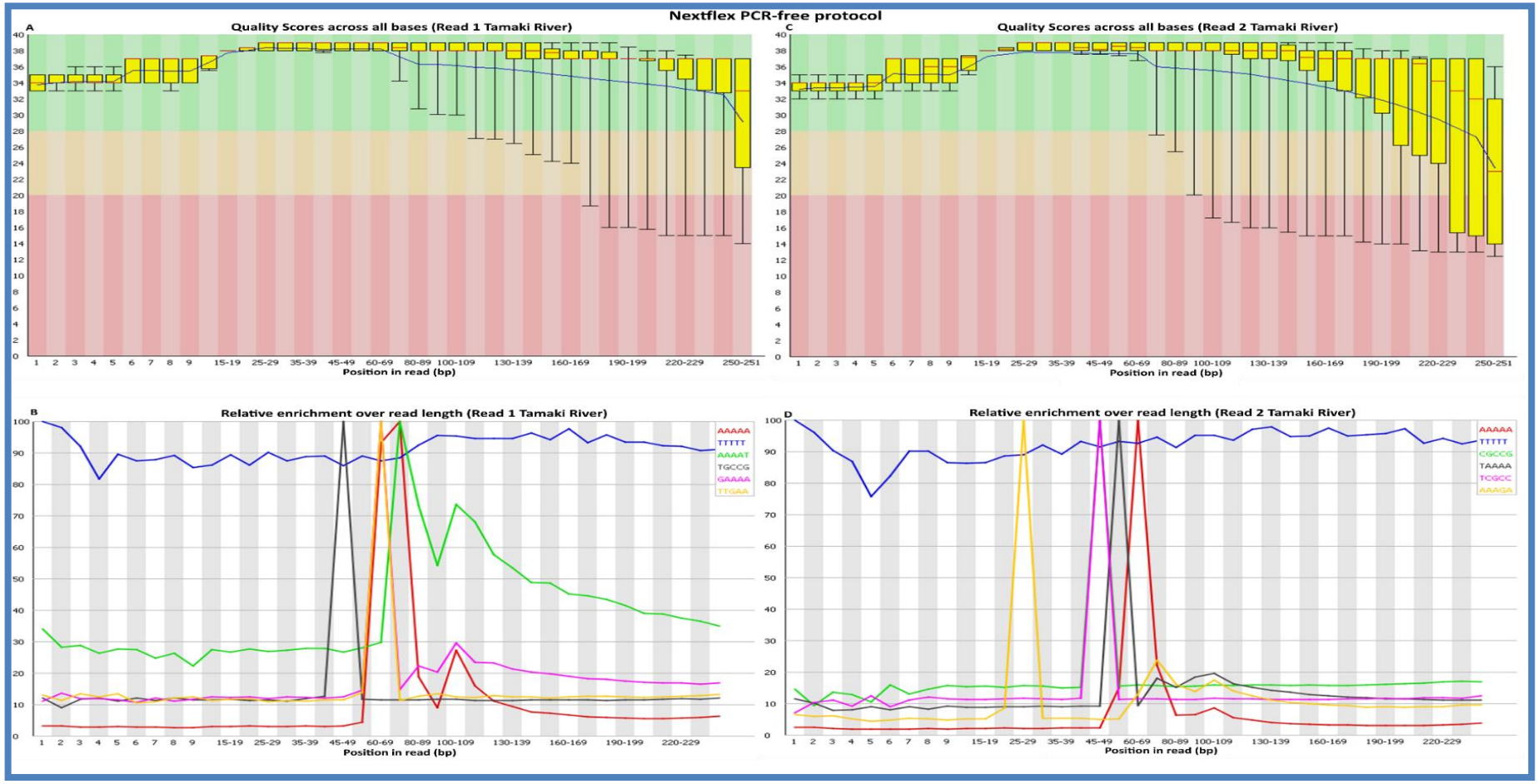


Figure 30 – A) Sequence quality for read 1 was high with almost all sequences exhibiting a phred quality score above 28 and a 0.01% error rate. C) Sequence quality for read 2 was good before 200 bases and reduced to a lower phred score of 20 after 230 bases. B) The presence of 5-mer repetitive sequences likely due to primer adapter sequences or possibly dimer contamination. These kmers were also present in read 2 (D).

Data from the Ion Xpress protocol

The Ion Xpress 400 bp sequencing kit generated a single read length of 400 bp with 5,428,136 million raw single-end reads (Table 10). The preliminary FastQC report showed that the majority of the Ion-Torrent data had a phred score of more than 25 only for less than 250 bp sequences in which the report indicated that the read length of the majority of the sequences was only between 150 and 250 bp (instead of the expected of 400 bp long) with a GC content of 58% and sequence duplication level of 11.06% (Figure 31). There were no overrepresented sequences but we encountered a high peak of repetitive K-mer sequences from position 250 bp to 400 bp that indicated either a mixture of true repetitive regions or poor quality, homopolymer repetitive sequences (Figure 31).

Base quality and Kmer content in the Ion-Xpress 400bp sequences for Tamaki River one litre grab sample

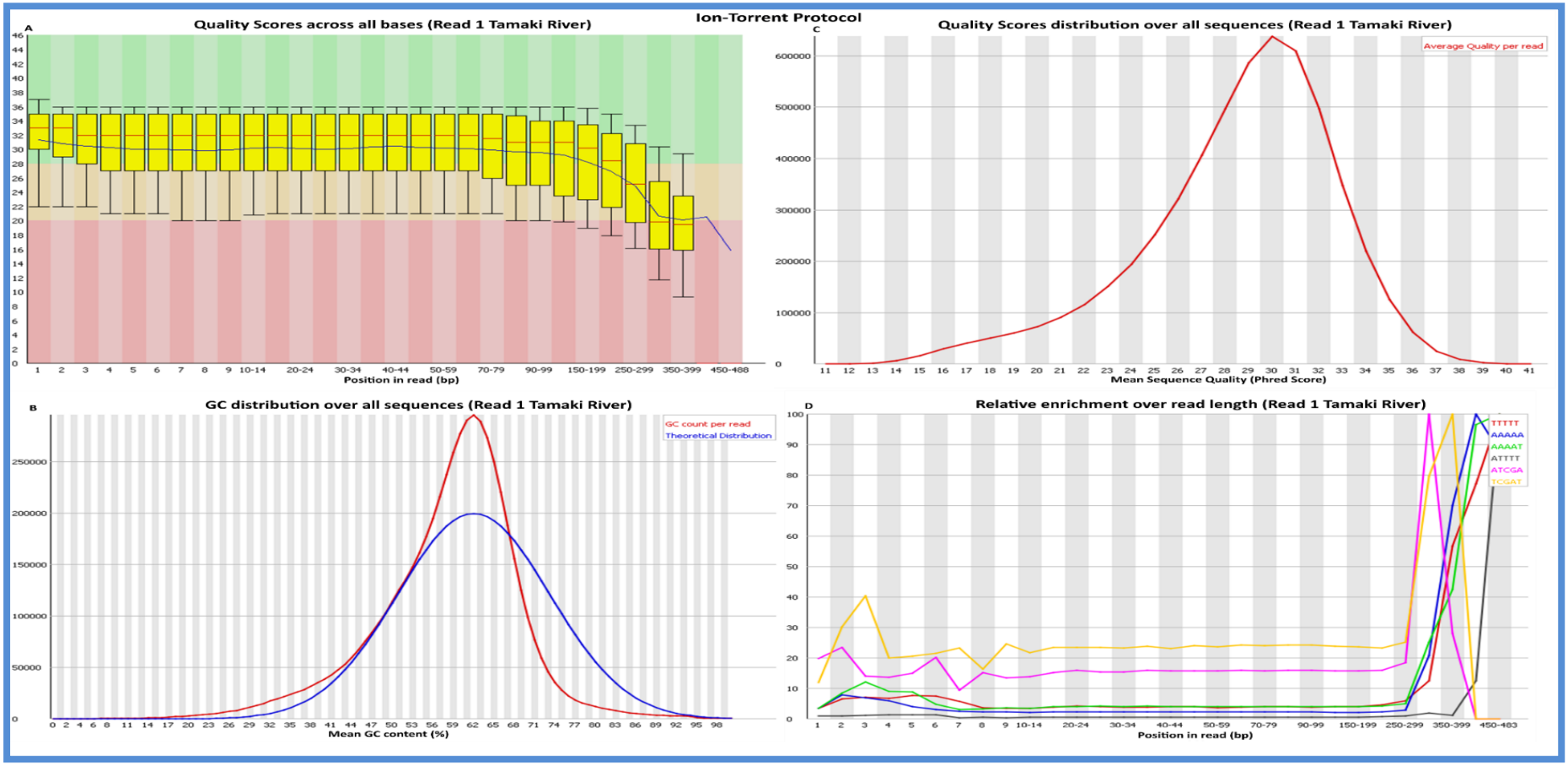


Figure 31 – Preliminary FastQC report for Ion-Torrent data indicating quality across length of all sequences. A) Sequences were generally of high quality up to position 200 bases before dropping to lower quality after 250 bases. B) The distribution of GC content over all sequences and peak at ~ 62% with at least 250,000 reads C) The distribution of quality scores for all DNA sequences indicating region of sequences with lowest error rate and we have at least 600,000 reads with a phred score above 30, D) This visualisation shows the presence of repetitive sequences among the reads located at 250 - 400 bp.

Comparison of FastQC quality distribution graphs for four different library preparations kits on Illumina and Ion-Torrent platforms

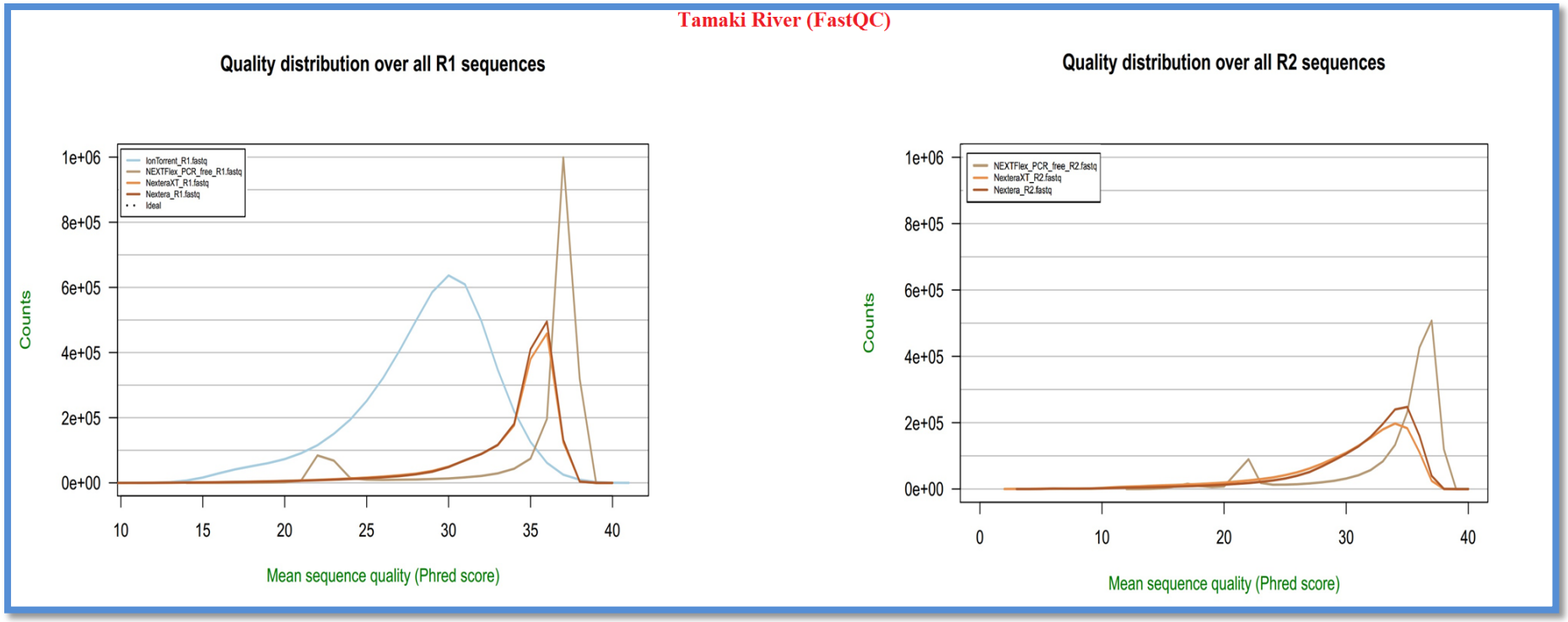


Figure 32 – A) Quality distribution graph for read 1 sequences shows that the majority of the raw metagenomic sequences sequenced using different Illumina library protocols were of good quality (Phred score of 30 and above) with 99.9% accurate base-calling. The Ion-Torrent data was of lower quality. B) Quality distribution graph for Illumina read 2 sequences. The graph again shows that the majority of the raw metagenomic sequences were of good quality with approximately 1% error rate of incorrect base-calling.

FastQC summary report for sequences obtained with different library protocols and platforms

Platform	Library Construction	Data Output (gigabases)	No of reads (Single reads)	Average Sequence Read Length (bp)	Error Rate (%)	GC Content (%)	Sequence Duplication Level (%)	Overrepresented Sequences	Kmers content (repetitive sequences)	QC Summary
MiSeq	Nextera	1.19	3,427,986 (1,713,993)	151	0.01	54 (Both R1 and R2)	5.67 (R1) 5.48 (R2)	No	AAAAA (R1 and R2)	Passed
MiSeq	Nextera-XT	1.15	3,306,580 (1,653,290)	151	0.2	54 (Both R1 and R2)	6.11 (R1) 5.7 (R2)	No	AAAAA (R1 and R2)	Passed
MiSeq	Nextflex PCR Free	2.12	3,880,968 (1,940,484)	251	0.02	53 (R1) 54 (R2)	11.17 (R1) 10.58 (R2)	Adapter Index 4 sequence (97% over 49bp)	AAAAA (R1 and R2)	Passed
Ion-Torrent	Ion Xpress 400 bp kit	2.01	5,428,136	400	0.02	58 (Read 1 only)	11.06 (Read 1 only)	No	TTTTT	Passed

Table 10 – Pre-processing (FastQC) quality assessment for sequences obtained using Nextera, Nextera-XT, NEXTFlex PCR free and Ion Xpress-400bp) sequencing protocols

3.6.2 Quality Assessment using SolexaQA

SolexaQA (Cox et al., 2010) was also used to quantify sequence data quality.

Data from Nextera and Nextera-XT protocols

The SolexaQA report indicated we had a total of 6,855,972 (Nextera) and 6,613,160 (Nextera-XT) sequences for both reads 1 and 2. To standardise the sequencing read length across all the platforms for a comparative study, the data were trimmed for adapter sequences using DynamicTrim (Cox et al., 2010), set to allow only a 1% error rate (p-value = 0.01). Low quality reads of less than 35 bp (mostly adapter) were then excluded from further analysis using LengthSort (Cox et al., 2010). After trimming, the SolexaQA report showed we had a mean segment read length of 126.5 bp and 102 bp and median segment length of 145 bp and 108 bp for both read 1 and read 2 respectively for sequences generated with the Nextera protocol, (Table 11). Meanwhile from the Nextera-XT protocol, we had a mean segment read length of 124.4 bp and 93.1 bp and median segment read length of 144 bp and 98 bp for read 1 and read 2 respectively (Table 11). The trimmed sequences were re-evaluated and the number of high quality paired-end sequences was reduced to 2,522,712 (73.59%) for Nextera and 2,204,324 (66.66%) for Nextera-XT (Table 11). For the sequence data obtained using the Nextera protocol we had a total of 336,464 (9.81%) single unpaired reads and 568,810 (16.5%) discarded reads and data from Nextera-XT protocol, we had 393,879 (11.9%) unpaired reads and 708,377 (21.4%) discarded reads (Table 11). Overall the sequence quality for both reads 1 and 2 from both sample preparation methods were similar and equally good (Figure 33 and 34).

Data from the NEXTFlex PCR-free library protocol

Next sequencing data generated from NEXTFlex PCR-free library protocol, the SolexaQA analysis yielded an output of 2.12Gb and 3,880,968 of raw sequences for 2 x 251 bp paired-end sequencing run (Table 11). Using default algorithm the sequence reads were trimmed with a default p-value of 0.05 (equivalent to quality score Q~13) to remove adapter dimers and further evaluated again via LengthSort which a program to separate good quality reads from lower quality reads. After trimming, the SolexaQA report indicated a mean segment read length of 135.6 bp (Read 1) and 101.7 bp (Read 2) and a median segment read length of 149 bp (Read 1) and 110 bp (Read 2) (Table 11). The report also showed the trimmed data

consisted of approximately 1,750,724 (90.2%) high quality paired sequences (above Phred=20), 141,052 (3.63%) single unpaired reads and 238,468 (6.14%) discarded sequences (Table 11). Generally the data output and quality from the NEXTFlex PCR-free method was comparable to that from the other library construction protocols (Figure 35).

SolexaQA generated reports for Nextera library protocol on a MiSeq

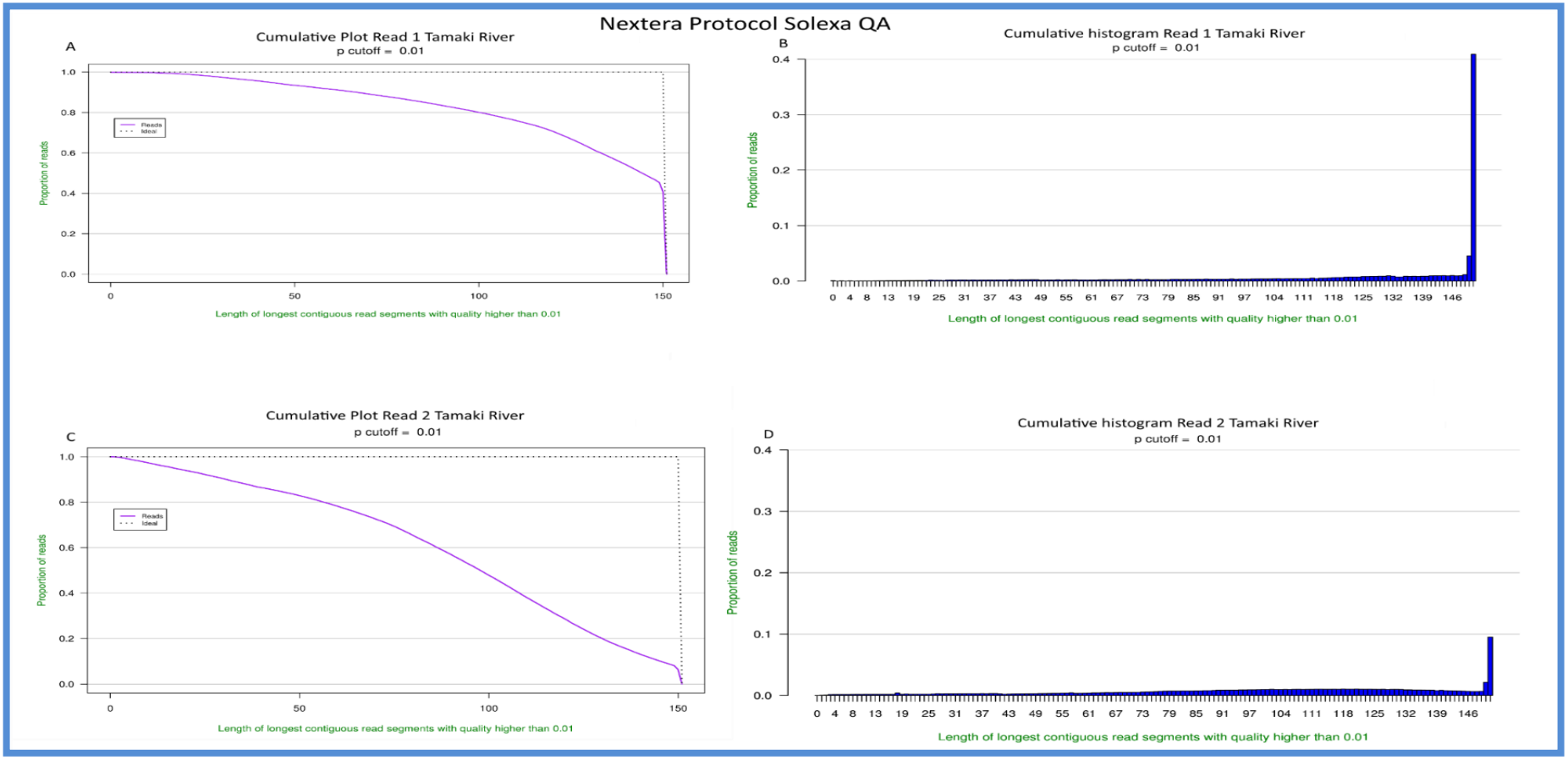


Figure 33 – SolexaQA cumulative plots and histograms showed that the majority proportion of our reads were of high quality. A) Almost 80% of our reads from read 1 have more than 100 bases and approximately 40% are at 150 bases, B) The higher quality reads are reflected on the histogram showing majority proportion of reads at 150 bases. Meanwhile we observe a drop in sequence quality for read 2 compared to read 1 where approximately 60% of reads are less than 100 bases and only 10% reads are at 150 bases (C) and this was further reflected on the histogram for read 2 (D).

SolexaQA generated reports for Nextera-XT library protocol on a MiSeq

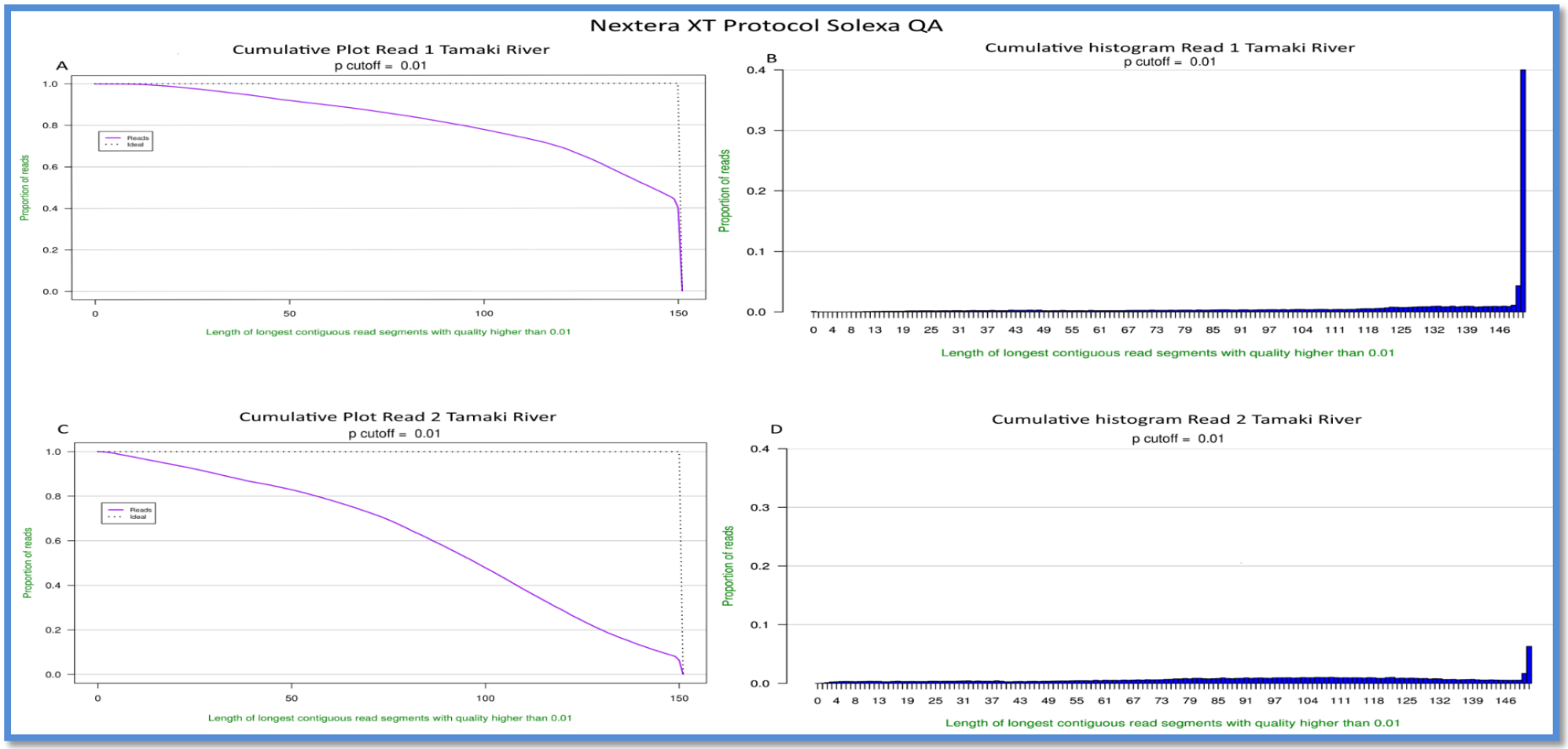


Figure 34 – Similar to the Nextera protocol, both cumulative plots and histograms for Nextera-XT protocol showed that the majority of our reads were of high quality. A) Approximately 75% of Nextera-XT reads (read 1) have more than 100 bases and less than 40% of the reads have 150 bases, (B) The 150 base reads can be seen on the histogram with 1% error rate. In comparison, the read 2 quality drop earlier compared to read 1 where approximately 50% of reads now are less than 100 bases and only less than 10% reads are at 150 base reads long (C) and again this was shown on histogram plot for read 2 (D).

SolexaQA reports generated with NEXTFlex PCR Free library protocol on a MiSeq.

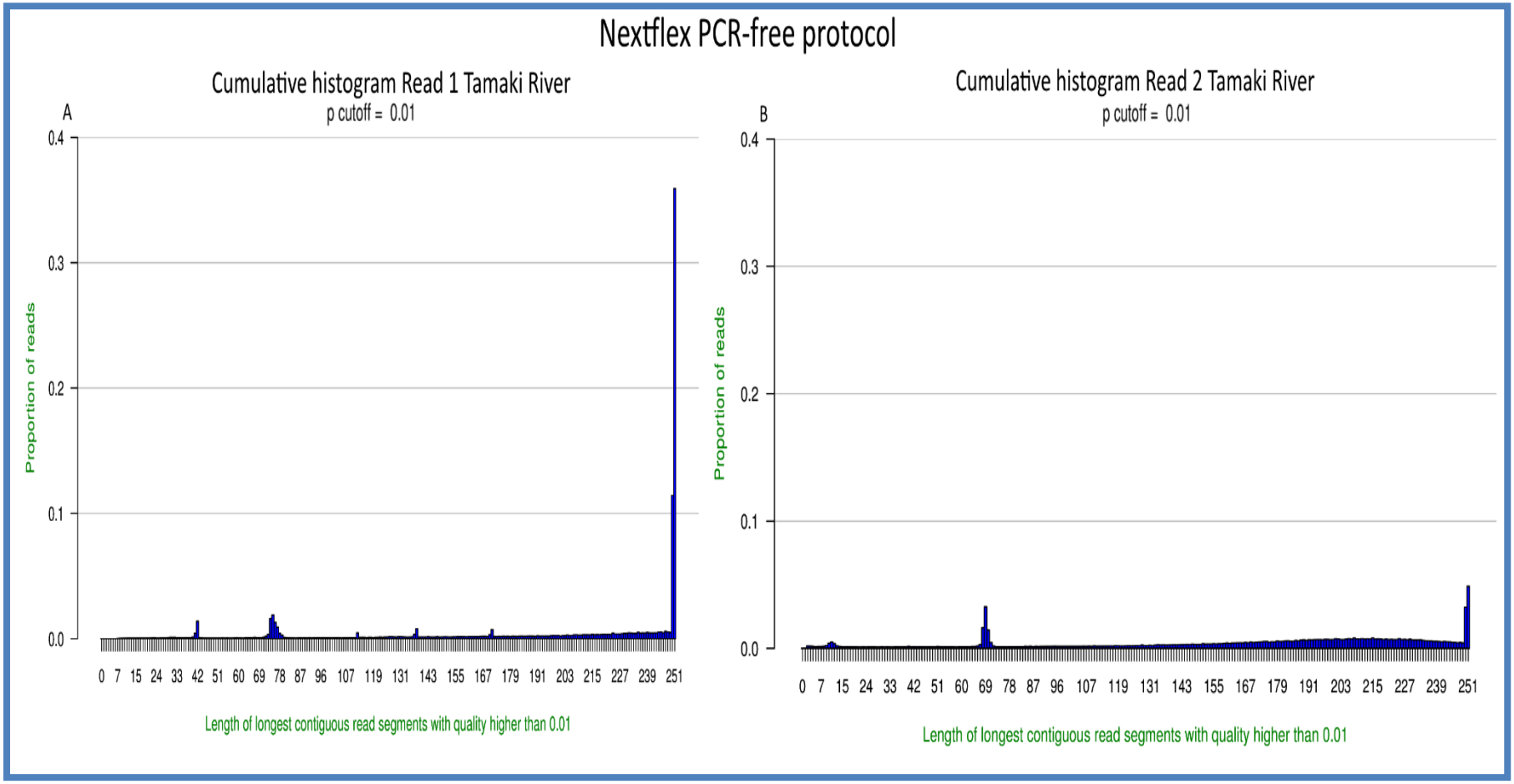


Figure 35 – Histogram plots which show that the quality of read 1 data was better than read 2 with most of the longest fragments (251 bp) being generated from read 1 with <0.01 error rate. A slight drop in sequence quality can be observed between reads with 60 and 80 bp long fragments. This is indicated by a small spike in both read 1 and read 2 histograms.

Data from the Ion-Xpress protocol

SolexaQA analysis showed that the longest contigs generated from the Ion-Xpress kit were 461bp. The total raw data output was ~2.01 Gb which was approximately 5,428,136 reads (Table 11). DynamicTrim was set to trim the adapter reads from the 461 bp long fragment reads with the parameter filter setting at 1% error rate (p-value = 0.01, phred = 20). This was also necessary, as the data showed a significant drop in quality after 216 bp where the quality phred score was at 10 and below. To trim the bad quality reads, the data was run through LengthSort software with the length parameter set to 75 bp or less. According to the cumulative plot majority of the trimmed sequences (75 bp or less) are still relatively lower in quality reads with ~ 2,229,013 (41.06%) better quality single reads (>phred = 20) compared to 3,199,123 (58.9%) discarded reads (Figure 36) (Table 11). In general, the data obtained from this run were not promising as we needed to trim from 461bp to 75bp or less to salvage readable sequences above a phred-score of 20 (41.06%). Such extensive trimming was necessary as the sequence quality was poor after 100 bases and resulted in a very low quality beyond 250bp. The run was not repeated due to budget constraints for this project.

SolexaQA reports generated with Ion-Xpress 400 bp sample preparation kit on an Ion-Torrent PGM

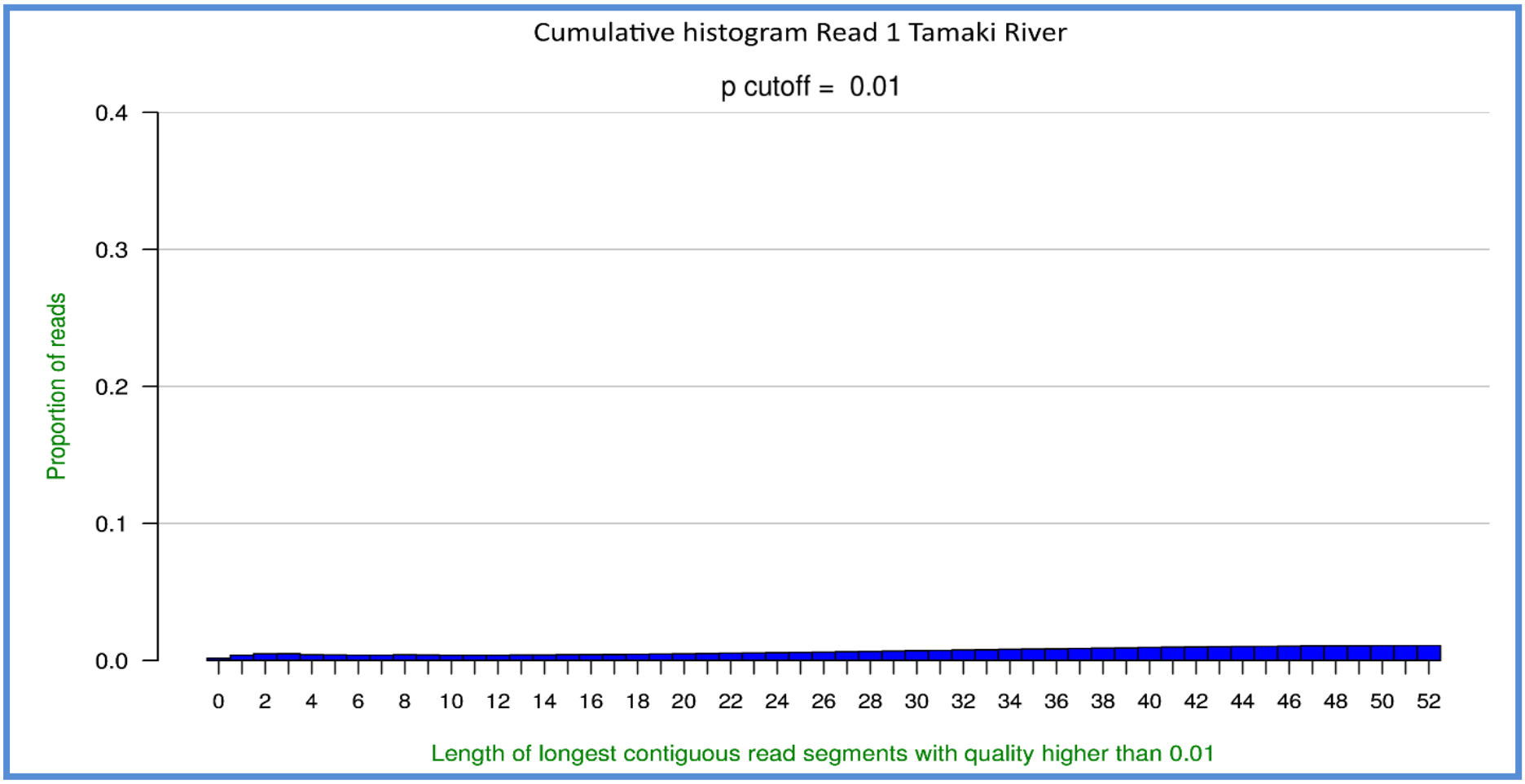


Figure 36 – Histograms with 1% error rate showing the length of the contiguous read data generated from the Ion-Torrent PGM after trimming. For trimming, the data was run through LengthSort set to 75bp with p-value of 0.01 and this gave 2,229,013 (41.06%) good quality single reads (phred score >20) and 3,199,123 (58.9%) discarded reads (Table 11). The figure shows only the relative proportion of sequencing reads after trimming with maximum fragments length at 52 bp

3.6.3 Summary of results from both QC software

In general, the FastQC java-based software for data quality assessment provided an efficient tool to check and evaluate the Tamaki River sequence data. This less intensive and low memory software provided a quick analysis of a relatively large set of metagenomic data. The assessment provided basic information on the number of processed raw reads: the number of raw clusters, % filtered reads, % of error rate per-base quality scores, % GC content, % duplication read (PCR artefacts) and a proportion of overrepresented sequences. The quality distribution curves and reports showed that the metagenomic reads were generally of good quality with the different library preparation protocols (Table 10, Figure 32).

SolexaQA generated similar quality reports for our data. SolexaQA was slower to run than FastQC however, it offered a more convenient package with DynamicTrim and LengthSort software in built, enabling us to easily gain a better understanding of data quality after trimming (Table 11, Figure 37). One advantage of SolexaQA is that it gave a more accurate indication of read quality with a default phred score of $Q_{13.}$ SolexaQA also required less computing resources (Del Fabbro et al., 2013).

Generally SolexaQA analysis showed that the read qualities and quantities (from both MiSeq and Ion-Torrent generated data) were not similar. That is, we observed a decrease in quality for Ion-Torrent Read 1 data in comparison to MiSeq Read 1 sequences after 200 bases (Figure 37). According to SolexaQA cumulative plot majority of Ion-torrent sequences have a read length of less than 200 bp with no fragments exceeding 250 bp or more (Figure 37). The majority of read 1 sequences from the MiSeq data had reads longer than 100 bp. The data from the NEXTFlex PCR free protocol on the MiSeq instrument yielded the highest proportion of reads with lengths greater than 250 bases.

Comparison of SolexaQA cumulative plot for all metagenomic data generated from MiSeq and Ion-Torrent PGM

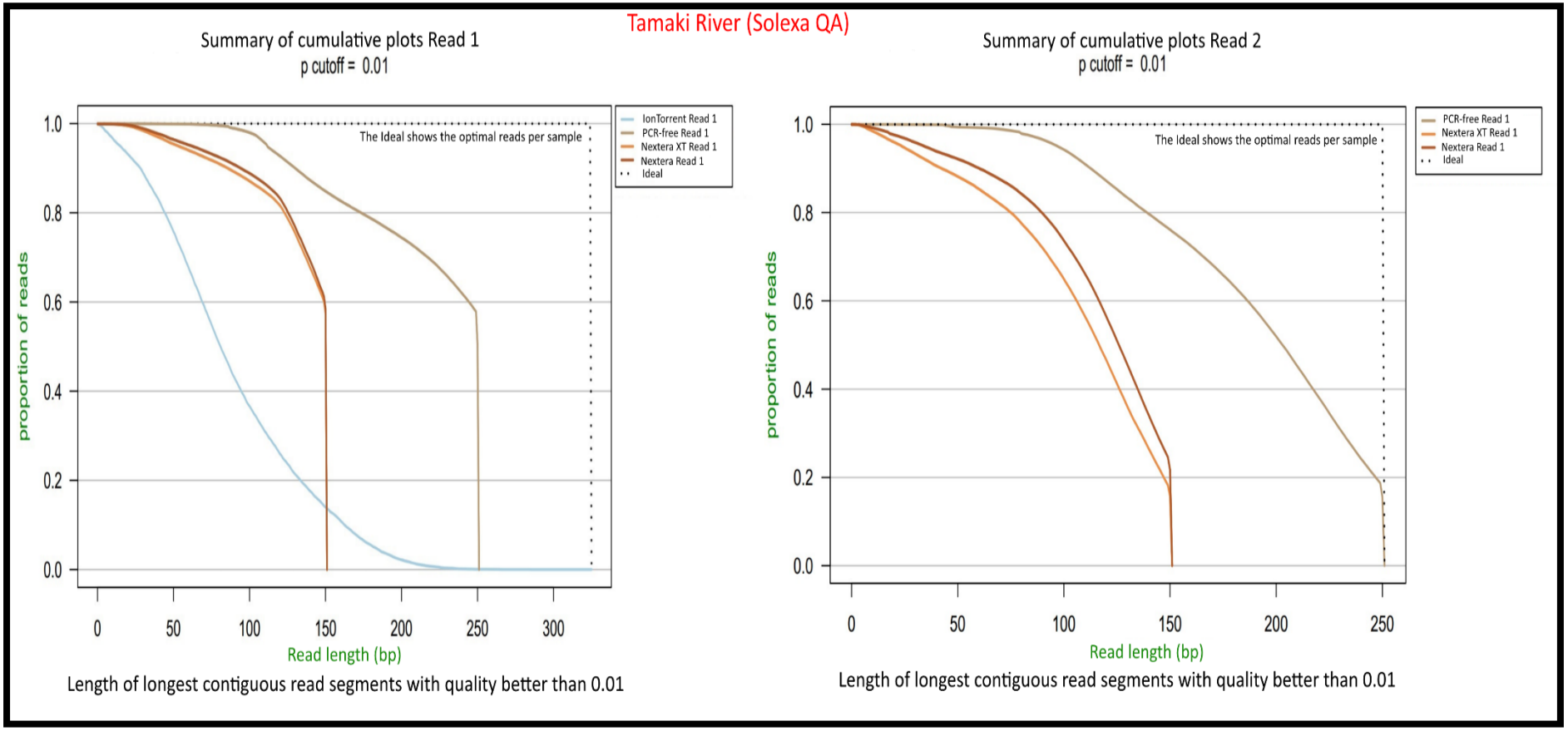


Figure 37 – Combined cumulative SolexaQA plots for metagenomic data obtained using different library preparation protocols. A) Summary for read 1 showed a comparison of sequences from two sequencing platforms and different library protocols utilised in this project. We observed a mixture of high and poor quality data across the library protocols and across the platforms (B) Cumulative plot for Read 2 data indicating that the majority of sequences (~50%) with Q_{20} scores were less than 100 bases in length. There is no data for Read 2 for Ion-Torrent because it was only a 400bp single read run since paired-end read chemistry is still not available.

Comparison of SolexaQA summary reports

Platform	Library Preparation Kit	Raw Data Output (Gigabases)	No of Raw Sequencing Reads	Sequence Read Length Chemistry (Bp)	Mean Segment Read Length (Bp)	Median Segment Read Length (Bp)
MiSeq	Nextera (Tamaki River)	1.19	3,427,986 (R1, R2)	151	126.5 (R1) 102.0 (R2)	145 (R1) 108 (R2)
MiSeq	Nextera-XT (Tamaki River)	1.14	3,880,968 (R1, R2)	151	124.4 (R1) 93.1 (R2)	144 (R1) 98 (R2)
MiSeq	NEXTFlex PCR Free (Tamaki River)	2.12	3,880,068 (R1, R2)	251	135.6 (R1) 101.7 (R2)	149 (R1) 110 (R2)
Ion-Torrent PGM	Ion Xpress 400 bp Kit (Tamaki River)	2.01	5,428,136 (R1, R2)	400	261 (R1) 152 (R2)	241 (R1) 130 (R2)

Table 11 – Summary of SolexaQA reports software for Nextera, Nextera-XT, NEXTFlex PCR free and lastly Ion Xpress-400bp data.

In our metagenomics dataset, we also observed some slight differences in sequencing coverage, read length, base composition (% GC) of the reads and the presence and resolution of homopolymers repeats between sequencing platforms on different chemistries. For example, the MiSeq sequencing coverage and % GC content from the Illumina Nextera protocol on the 2x150 PE run generated about 3 million reads (1.19 Gb) of sequencing data with a GC content of 54%. Meanwhile the Nextera-XT protocol on the 2 x 150bp PE run also generated approximately 3 million reads (1.15Gb) of sequencing data with a GC content of 54%. Both Nextera runs had a similar profile which indicates a sample consistency across the different library preparation protocols. Additionally, there were also slight variations in th GC content from approximately 53% on the MiSeq sequencer (PCR Free) to 58% on the Ion-Torrent PGM platform. Further investigation revealed the increase of GC content was due to the presence of homopolymers towards the end of the Ion-Torrent reads (>300 bp). A comparison of assembled contigs between Illumina and Ion-Torrent data showed differences in sequence duplication levels of 7.45% for Illumina reads and 11.06% for Ion-Torrent data (Table 10). Also additional investigation revealed that the Illumina homopolymer reads were biased towards A's over T's (Ion-Torrent) nucleotides even though both originated from the same sample. This was confirmed by the high occurrence of A's and T's pentamers within the sequences (Table 10). PCR biases in our Nextera libraries appeared to be minimal. This can be concluded from the similar GC contents measured in PCR and PCR-free libraries.

We also compared the generated average read length and quality from both the MiSeq and Ion-Torrent sequencing platforms. We expected the Ion-Torrent to produce the longest average mean reads of at least 400bp, but due to poor read quality the longest read fragments were only 261bp after trimming. Thus the data lengths were comparable with the Illumina generated MiSeq shorter reads fragments i.e. average read length of 2x250bp (NEXTFlex PCR-free) and 2x150bp (both Nextera) at 237.3bp and 223bp respectively.

3.7 Secondary data analysis

Secondary analyses comprised of two phases: comparative (taxonomic) analysis and functional annotation. For the computational analyses we used PAUDA and MEGAN software (version 5). The data from the Tamaki River sample was evaluated for taxonomic composition and annotated using gene and protein prediction tools based on the SEED and KEGG-orthology databases. The analyses provided an overview of the functional characteristics of the river water community.

3.7.1 Taxonomy classification of metagenomics reads

The processed reads were matched using PAUDA against a local NCBI Protein database. The BLASTX-like outputs (*.rma* files) were then analysed using the software MEGAN5 (version 5.71) using the LCA algorithm which binned matched sequences under the most appropriate taxonomy node. The microbial profiles indicated that a large proportion of sequence reads were from bacterial phyla Actinobacteria, Bacterioidetes, and Proteobacteria (mostly the classes alpha-, beta- and gamma-Proteobacteria).

All Illumina data sets (from Nextera and Nextera-XT and NEXTFlex PCR-free protocols) showed that *Pseudomonas fluorescens* was the most common species in all cases with the Nextera protocol producing about 525,383 matched reads, Nextera-XT produced 482,640 matched reads, and NEXTFlex PCR-free produced 680,861 matched reads (Figure 38-44). The next most abundant species was surprisingly the bacterium *Yersinia enterocolitica,* which previously has been reported in New Zealand in river water contaminated with the blood and run-off from pigs, cattle and deer (Bottone, 1999). This species was identified by 282,150 Nextera matched reads, 261,582 Nextera-XT matched reads and 481,131 NEXTFlex PCR-free with matched reads, (Figure 38-40). The Ion-Torrent instrument also returned a similar profile to the MiSeq data sets, with the species *Pseudomonas fluorescens* (981,825 reads) and *Yersinia enterocolitica* (334,391 reads) having the highest number of matching reads (Figure 41-42).

Generally, we observed similar taxonomic profiles for the Ion-Torrent and Illumina MiSeq generated metagenomics datasets (Figure 43-44). Bacteria from the family of Enterobacteriaceae, Comamonadaceae, Flavobacteriaceae, Oxalabacteraceae, Burkholderiaeae, Aeromonadaceae, Shewanellaceae and Moraxellaceae were similarly

represented in both the Illumina sequencing chemistry and Ion-Torrent datasets. That is similar proportions of commonly occurring bacterial species were observed in data from both instruments.

Twenty most represented bacterial 'genera' found in 1 litre 'grab' water sample from Tamaki River (MiSeq Data Only)

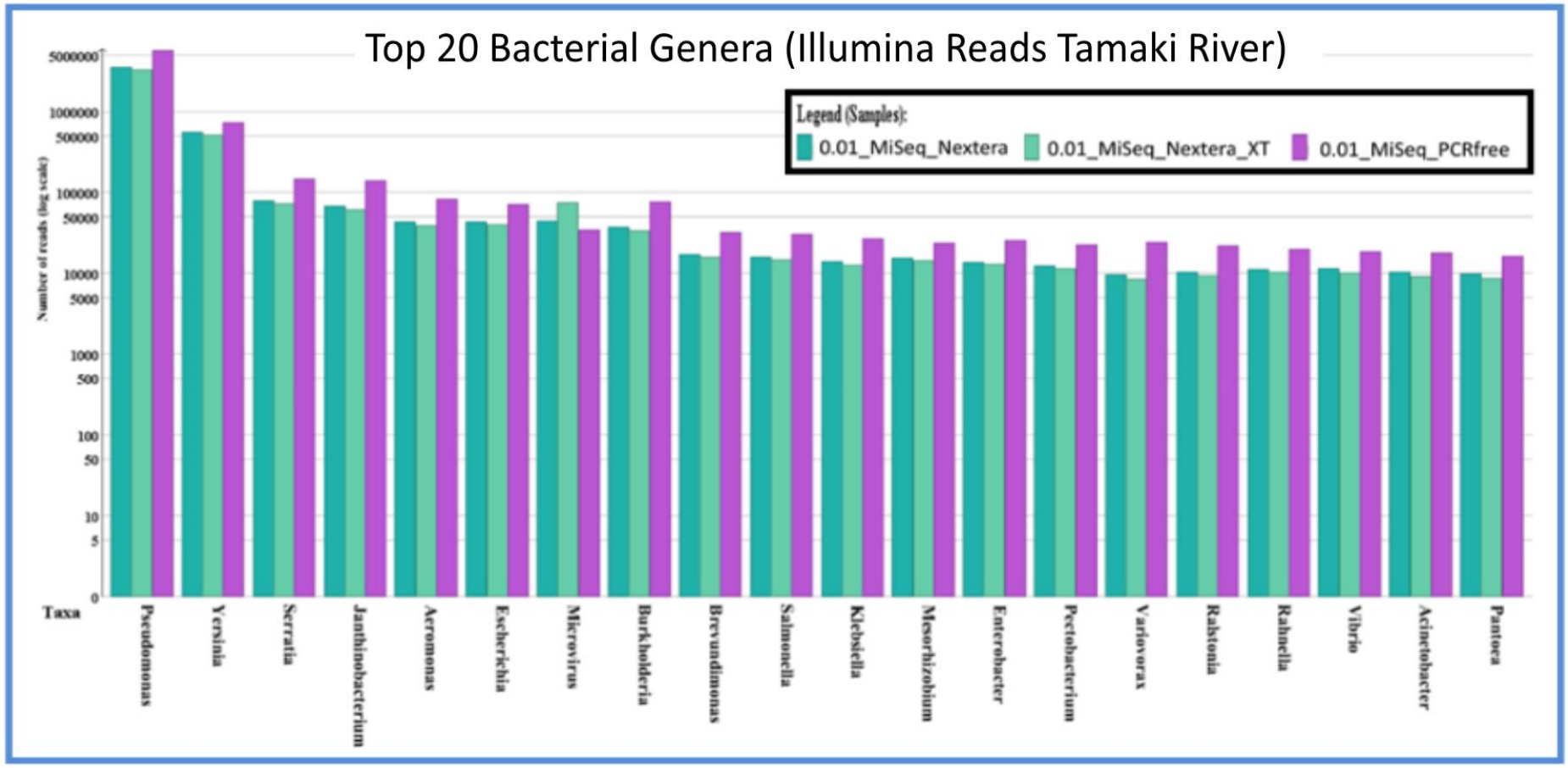


Figure 38 – Twenty most represented bacterial genera in the Tamaki River sample (number of reads indicated via log scale algorithm). The top three bacterial genera were *Pseudomonas*, *Yersinia* and *Serratia*. The presence of *E. coli* in our Tamaki River sample (via colorimetric result, see result section 3.1) was also consistent with our NGS data as the genus *Escherichia* was within the top 10 bacterial genera.

Top 40 bacterial 'species' found in 1 litre 'grab' water sample from Tamaki River (MiSeq Data Only)

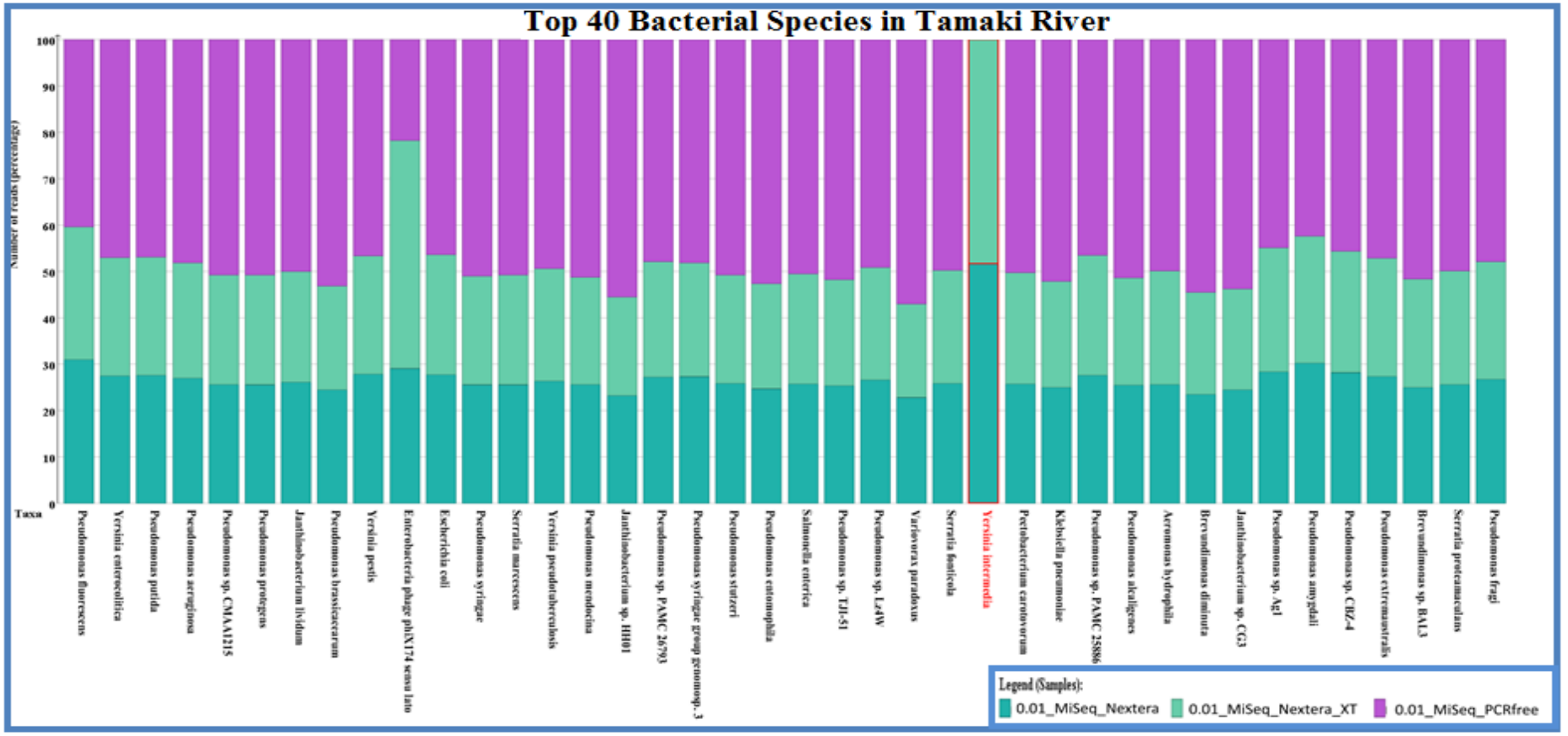
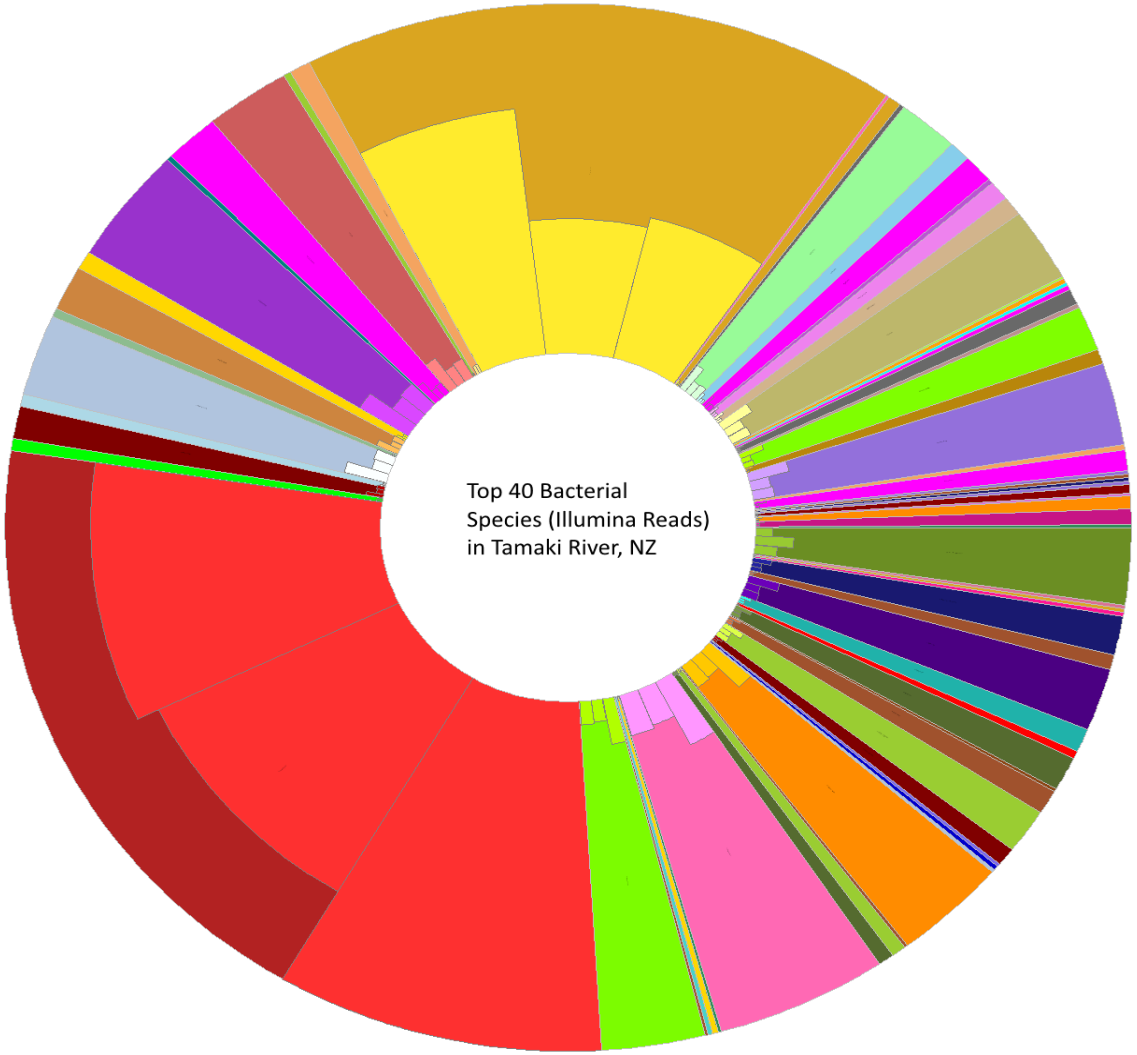


Figure 39 – Forty most represented species (99.5% hit/0.01 cut-off point) common to three different libraries (Nextera, Nextera-XT and NEXTFlex PCR-free). We observed two of the most abundant *Pseudomonas fluorescens and Yersinia enterocolitica* across all preparation methods. *E.coli* was observed to be the 11th most abundant species which is consistent with our calorimetric result (see result section 3.1). The number of sequencing reads for NEXTFlex PCR-free, Nextera and Nextera-XT protocols were similar.

Pie Chart showing the relative proportion of taxa identified in the Illumina libraries.



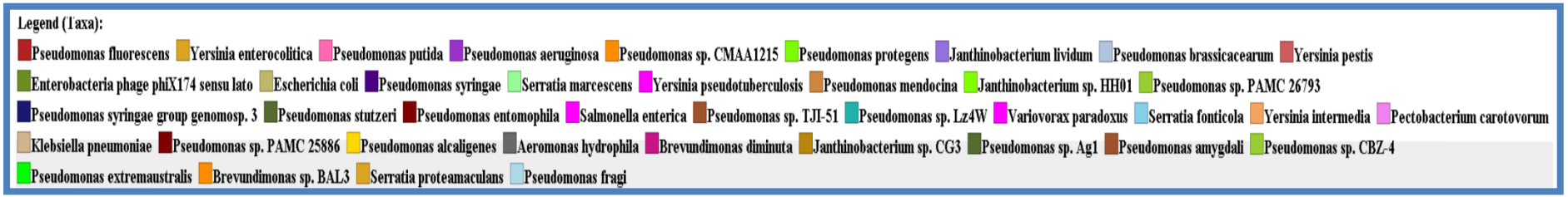


Figure 40 – Relative proportions of identified taxa generated from the Illumina MiSeq instrument (Nextera, Nextera-XT and NEXTFlex PCR-free protocols)

Forty most represented bacterial 'species' found in 1 litre 'grab' water sample from the Tamaki River (Ion-Torrent data only)

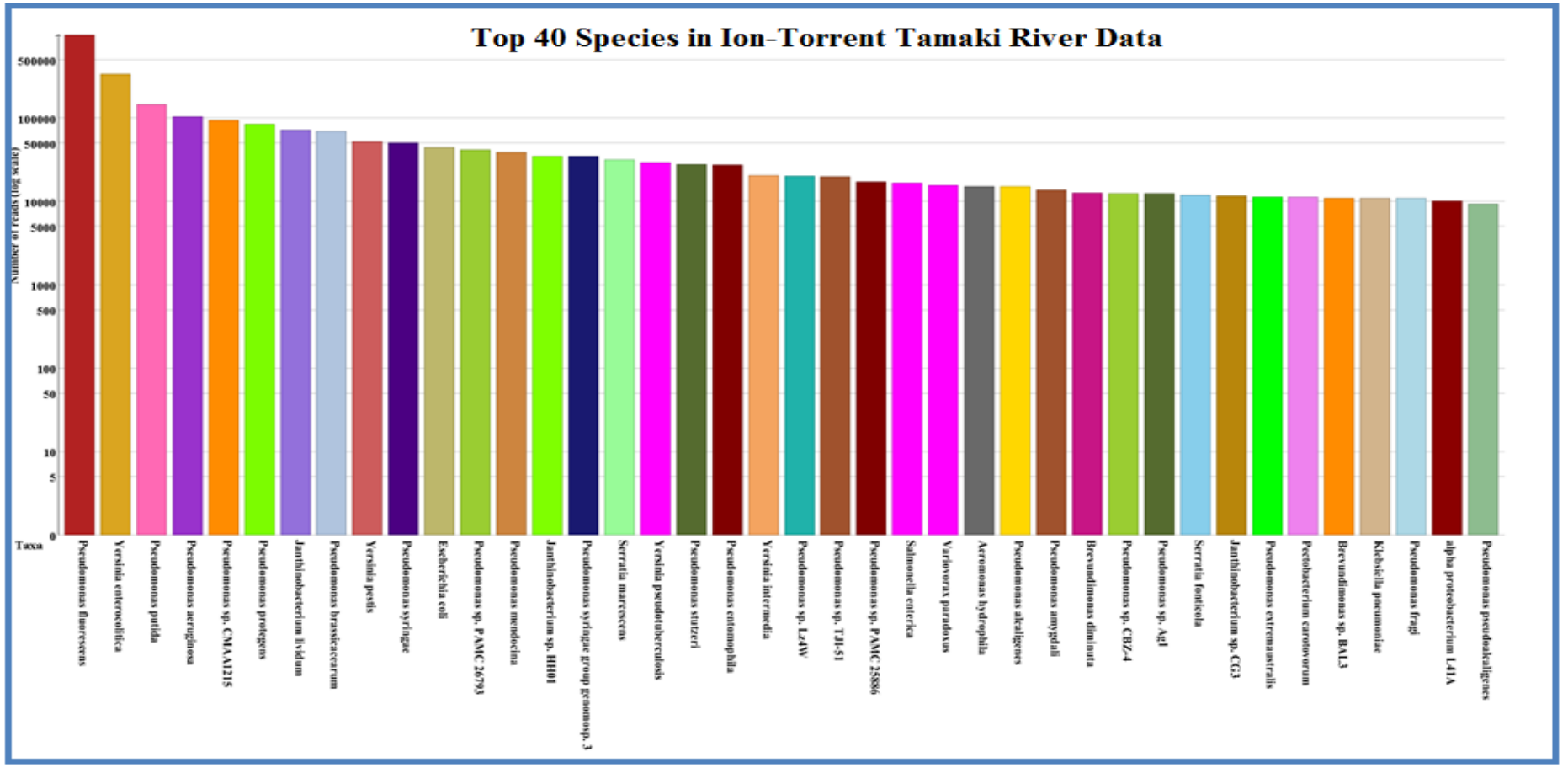
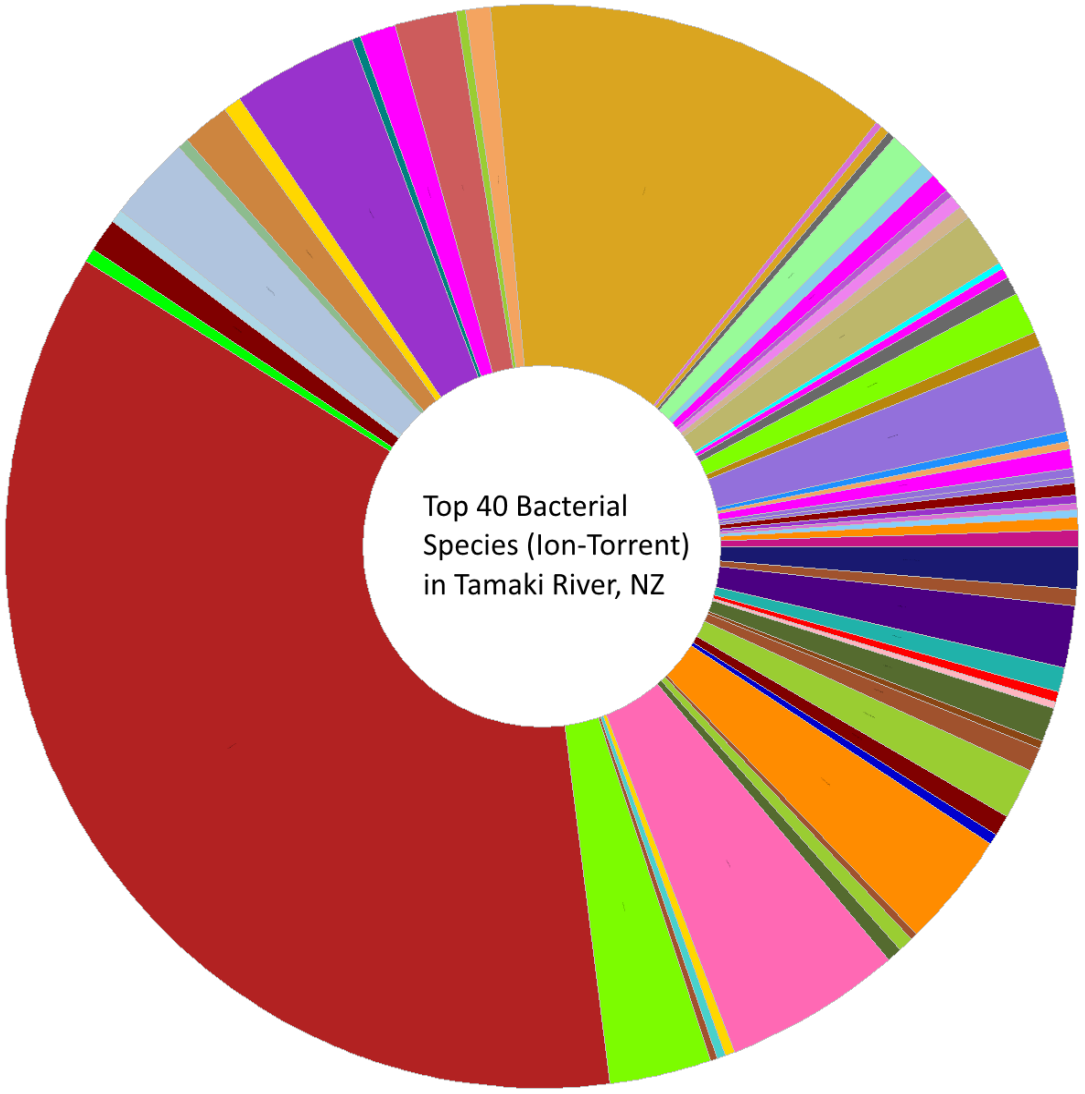


Figure 41 – Forty most abundant genera identified in the Ion-Torrent library. The profile is similar to that observed with the Illumina libraries in that *Pseudomonas fluorescens* (981,825 reads) and *Yersinia enterocolitica* (334,391 reads) species were most abundant, followed closely by *Pseudomonas putida* (145,957 reads), *Pseudomonas aeruginosa* (103,367 reads) and *Pseudomonas sp, CMAA1215* (93,998 reads). The low abundance of the bacterium *Esherichia coli* (43,901 reads) was also evident.

Pie Chart showing the relative proportion of taxa identified in the Ion-Torrent library.



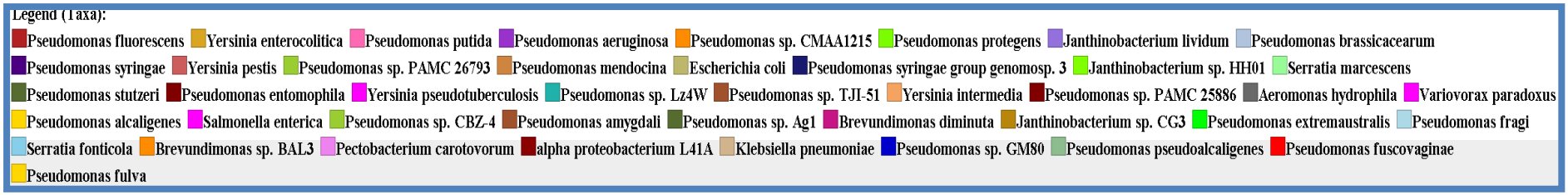


Figure 42 - Pie chart indicating relative proportion of the most abundant genera in the Ion-Torrent library. According to the analysis, *Pseudomonas fluorescens* (981,825 reads) made up more than 30% of the total bacterial population found in Tamaki River Water sample.

Whole bacterial population found in Tamaki River

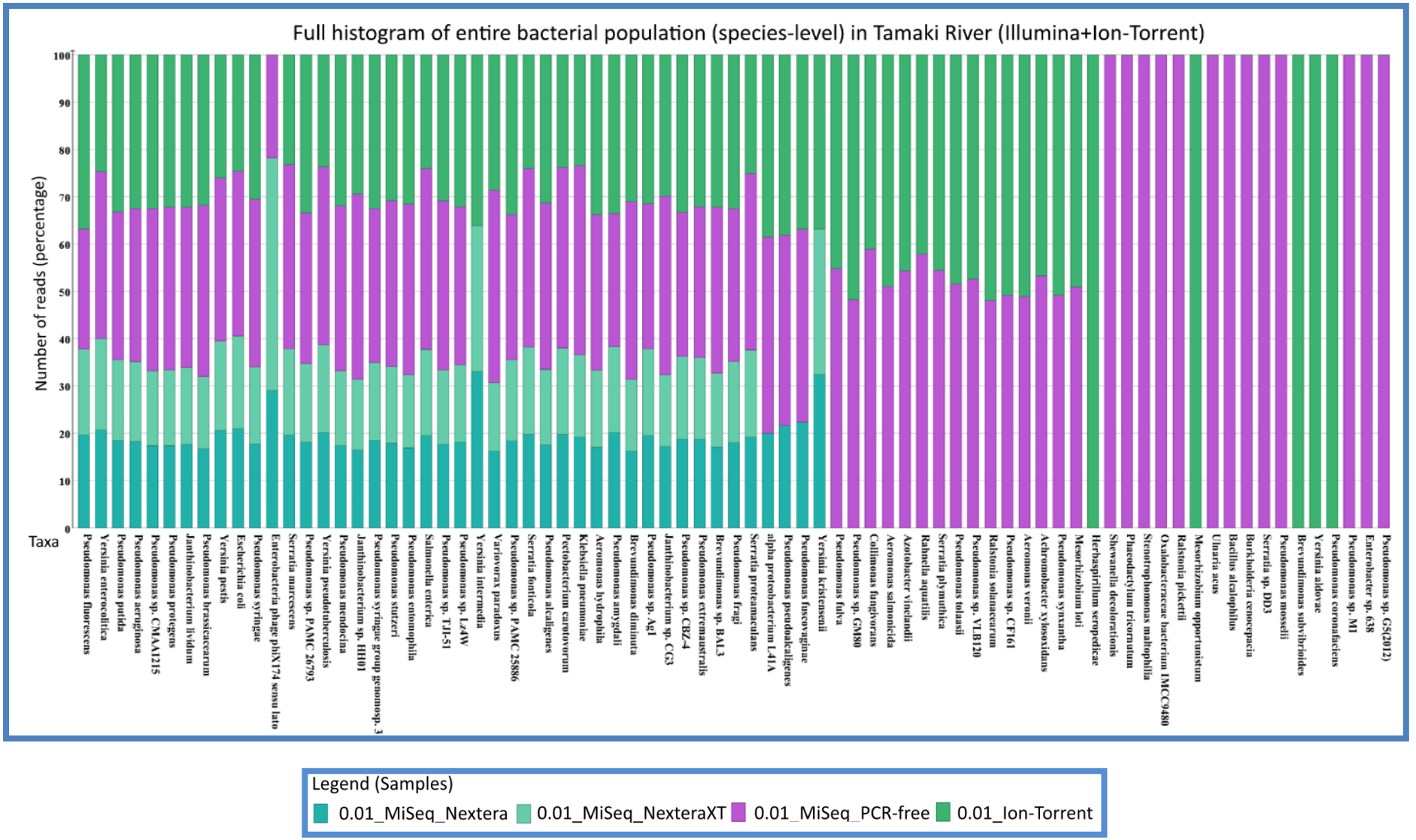


Figure 43 – Histogram of entire microbial profile (instead of just top 20 species) found in the Tamaki River obtained from all metagenomic datasets generated from different library preparation protocols and sequencing platforms. The bacterial composition for the 40 most abundant species was similar for different NGS protocols. The PCR-free protocol which used a longer read length (250 PE) sequencing chemistry had better sensitivity in detecting more species compared to other protocols. Five species: *Herbaspirillum seropedicae*, *Mesorhizobium* opportunistum, *Brevundimonas subvibrioides*, *Yersinia aldovae* and *Pseudomonas coronafaciens* were only present in the Ion-Torrent data. In addition, the bacterial '*phiX174'* was absent in the Ion-Torrent data which require further explanation.

Summary of entire bacterial 'species' using 'word cloud' presence in Tamaki River from both MiSeq and Ion-Torrent Platforms dataset

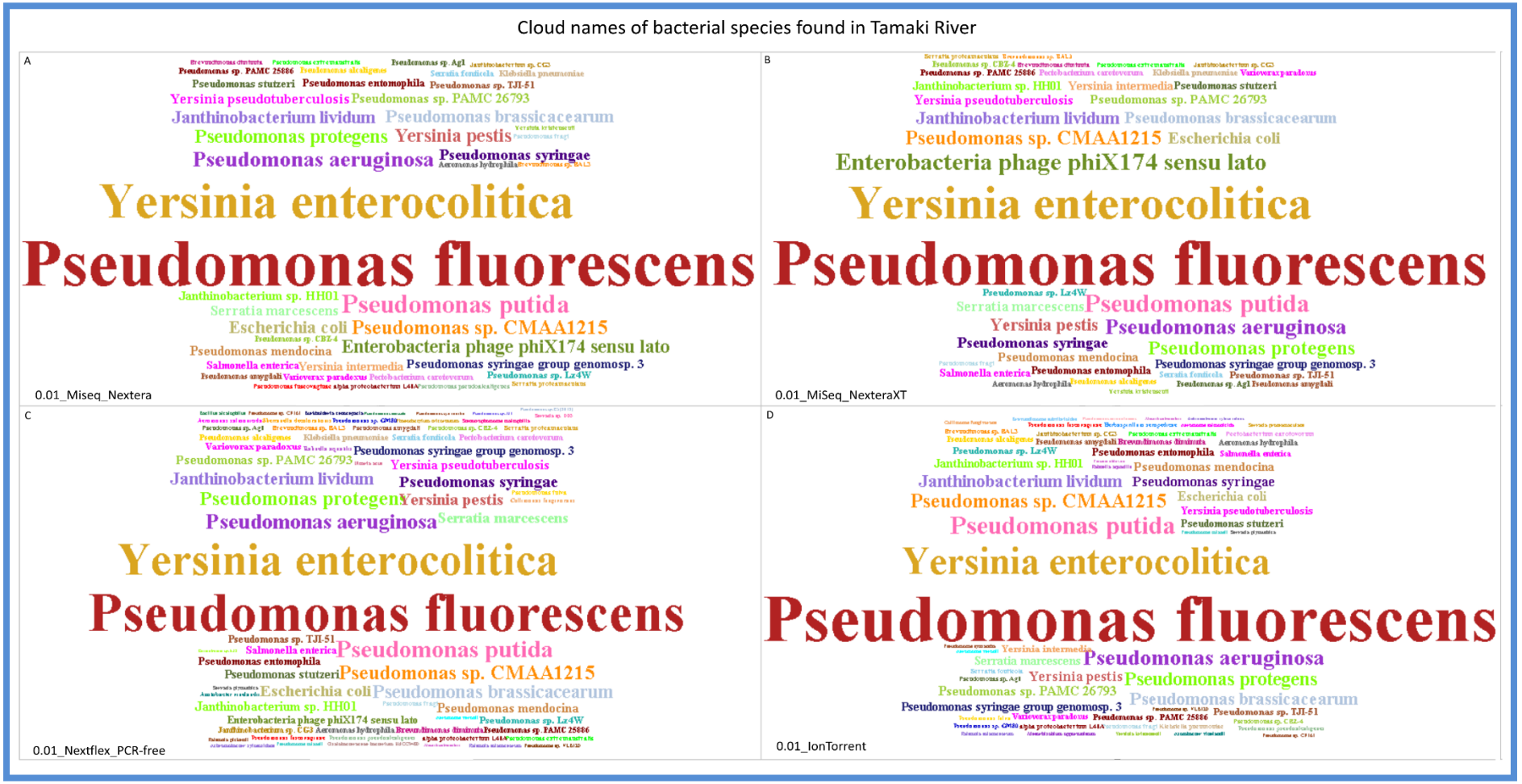


Figure 44 – A word cloud for all the metagenomic datasets originated from the MEGAN5taxonomy profiles (A: Nextera, B: Nextera-XT, C: PCR-free and D: Ion-Torrent) showing the most abundant bacterial species in the Tamaki River sample.

3.7.2 Functional analysis of metagenomic data using SEED and KEGG

For functional analysis of our metagenomic data, we utilised MEGAN 5 (version 5.7.1) to annotate and assign sequences to nodes using SEED and KEGG classifications.

3.7.2.1 SEED hierarchy with MEGAN5

Sequences from the Illumina prepared libraries (Dataset 1 to 3: Nextera, Nextera-XT, PCR-free protocols) and the Ion-Torrent library (Dataset 4) were loaded into MEGAN for SEED classification. Following annotation, the results were tabulated and presented as predicted "functional metabolic groups" for comparison of metagenomic profiles between libraries.

Our SEED analysis for both Illumina and Ion-Torrent reads gave similar results. In our analysis, the Illumina reads (Dataset 1= 7,777,570 (Nextera), Dataset 2 = 7,320,313 (Nextera-XT), Dataset 3 = 11,117,712 (PCR Free) and Dataset 4 = 12,299,859 (Ion-Torrent PGM) showed no obvious difference in terms of metabolic profiles (Table 12). Table 12 indicates that the highest number of reads were for carbohydrate synthesis (484,366), amino acids and derivatives (436,880 reads), protein metabolism (254,009 reads), DNA and RNA metabolism (218,678 and 144,669 reads), virulence, disease and defence metabolism (175,504 reads) (Table 12) (Figure 45). The different sequencing protocols did not uncover any significant differences in the metabolic gene content. Meanwhile, the number of unclassified sequences from both Illumina reads (Dataset 1 = 7,232,485, Dataset 2 = 6,818,347, Dataset 3 = 10,108,878, and Ion-Torrent Dataset 4 = 11,387,920) (Table 12) were also similar. There were many unassigned clusters of sequences in our SEED analysis and this is probably due to numerous reasons. The first issue is the repetitive sequences in our data coming from homopolymers in the genome. Second, given that we are dealing with aquatic metagenomes where the SEED subsystem may not include sequences from these new genomes in the current database, it is possible that these sequences are not represented in the database. Lastly, the stringency of the LCA algorithm may have also contributed to the large number of unassigned reads during BLAST analyses for sequence comparison and classification (Huson et al., 2007).

SEED classification heat map analysis for all metagenomic datasets

SEED Classification	Total reads for MiSeq Nextera (Data 1)	Total reads for MiSeq Nextera XT (Data 2)	Total reads for MiSeq Nextflex PCR-free (Data 3)	Total reads for Ion-Torrent (Data 4)
Not assigned	7232485	6818347	10108878	1387923
Carbohydrates	90192	82832	168207	143135
Amino Acids and Derivatives	79764	73523	140491	135102
Protein Metabolism	48391	44770	81642	78203
DNA Metabolism	40812	37448	74647	65771
Cofactors, Vitamins, Prosthetic Groups, Pigments	36018	32954	69919	60988
Cell Wall and Capsule	31465	28797	59833	50797
Respiration	28151	26576	49788	48153
Virulence, Disease and Defense	28143	25424	53295	48823
RNA Metabolism	27399	25276	46616	45378
Stress Response	24842	23104	45682	42066
Membrane Transport	24413	22887	48669	42317
Clustering-based subsystems	22256	20786	38846	36975
Nucleosides and Nucleotides	22003	20027	39237	36161
Iron acquisition and metabolism	21906	20049	46658	41504
Motility and Chemotaxis	19675	17715	34471	34605
Regulation and Cell signaling	17689	16391	34976	29571
Fatty Acids, Lipids, and Isoprenoids	15278	14257	30123	26593
Cell Division and Cell Cycle	14189	13317	23631	24425
Nitrogen Metabolism	12755	12145	24396	21774
Metabolism of Aromatic Compounds	10647	9938	21931	19191
Phosphorus Metabolism	9294	8273	18687	16078
Sulfur Metabolism	8384	7513	15958	14061
Miscellaneous	7491	6763	12844	12511
Potassium metabolism	5900	5283	10153	9959
Phages, Prophages, Transposable elements, Plasmids	2750	2504	4715	4159
Secondary Metabolism	2559	2525	4868	4277
Virulence	894	884	2593	1419
Phages, Prophages, Transposable elements	462	412	933	680
Photosynthesis	199	168	594	207
Dormancy and Sporulation	6	5	24	4
No hits	0	0	1	0

Table 12 – SEED subsystems classification on four metagenomic dataset: MiSeq Nextera (dataset 1), MiSeq Nextera-XT (dataset 2), MiSeq NEXTFlex PCR Free (dataset 3) and Ion-Torrent PGM (dataset 4). The number of assigned reads were filtered under standard correlation of 0.01 error rate to each functional biological nodes. Please note highlighted area shows large proportion of the NGS reads were binned to ‘unknown’ or ‘unassigned’ due to ambiguous nucleotide base-calling i.e. homopolymeric regions with many repetitive sequences.

Most represented SEED functional assignment categories

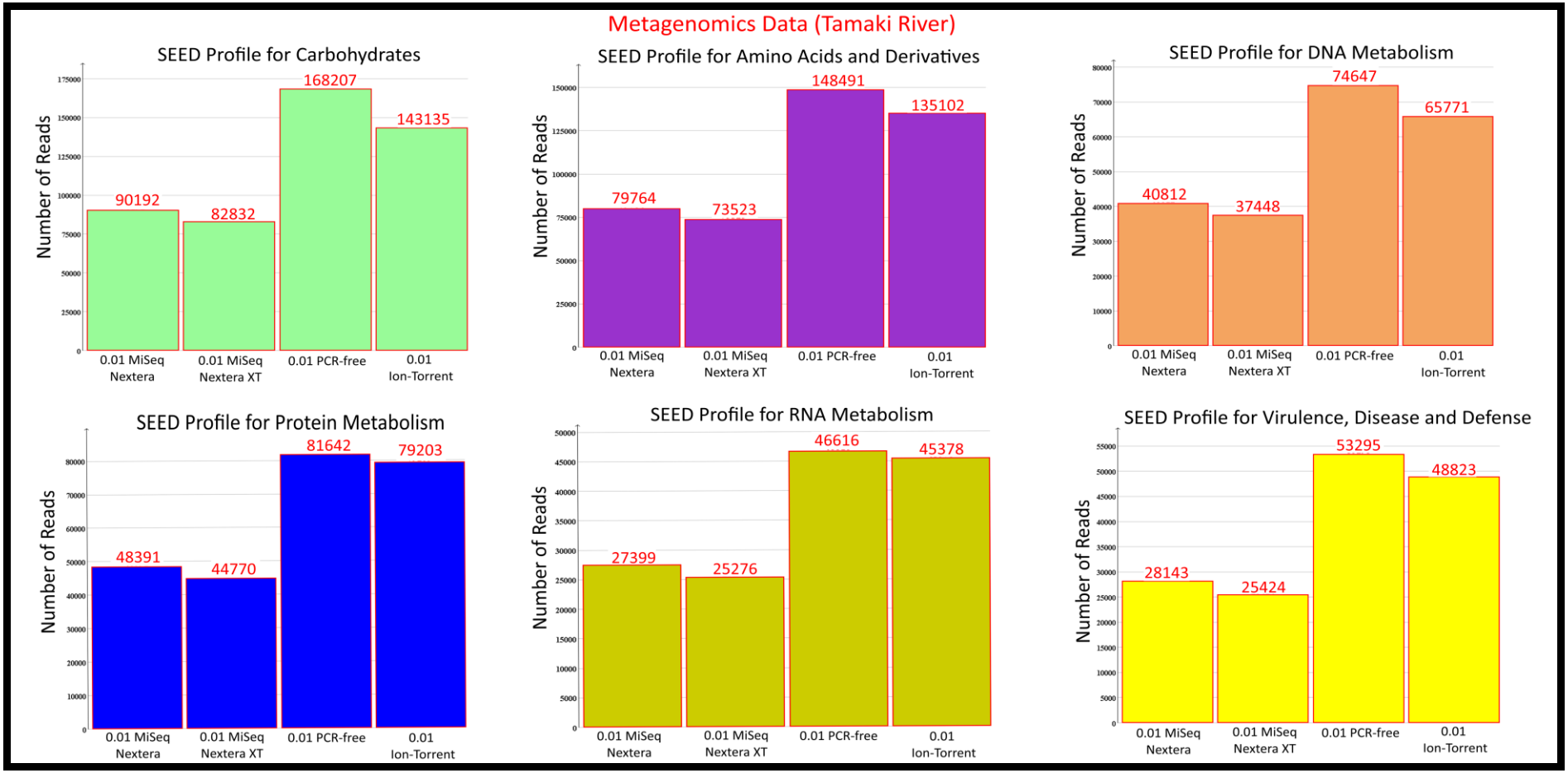


Figure 45 – Summary of SEED subsystems analysis showing different library sequences assigned to different categories of biological niche. These categories were carbohydrate synthesis, amino acid and derivatives synthesis, protein, DNA and RNA metabolism along with virulence factors and associated disease. There was no normalisation of the sequences for this data analyses.

3.7.2.2 KEGG pathway with MEGAN5

The sequence data were also analysed using the KEGG pathway (Kyoto Encyclopedia of Genes and Genomes) orthology system. KEGG is a collection of orthologous gene groups assigned to functional roles and biological pathways. The KEGG database has primarily been used to investigate the properties of biological systems within an environmental sample, incorporating genomic, chemical and functional information analysis (Kanehisa et al., 2008).

The annotated protein sequences were mapped to the KEGG BRITE functional module which comprises many BRITE hierarchy files used to identify the biological function of genes homologous to those in the translated protein-DNA NCBI database. To estimate the biological roles of each sequenced read, translated protein sequences were blasted (BLASTX) against an NCBI database with a cut-off parameter LCA score of 50 (default; 50% identification of Lowest Common Ancestor (LCA) score) using MEGAN5 software to generate an EC (enzyme commission) number for functional annotation assessment. The number of successful Illumina reads assigned to KO numbers (KEGG Orthology identification tag number) were 1,308,999 (16.56% match) for the Nextera generated data, 1,208,043 (16.24% match) for the Nextera-XT data and 2,040,042 (18.02% match) for the PCR-free data (Table 13). The matching percentage reads were assigned KEGG annotations. Meanwhile the Ion-Torrent platform generated 2,160,134 annotated reads. Almost 17.26% of these could be assigned to a KO number within the KEGG molecular network (Table 13).

Statistical summary of KEGG annotation

KEGG Pathway Analysis	MiSeq Nextera (Total assigned KEGG reads= 7905887)	MiSeq Nextera XT (Total assigned KEGG reads=7438982)	MiSeq Nextflex PCR-free (Total assigned KEGG reads=11315910)	Ion-Torrent (Total assigned KEGG reads=12515432)
Not assigned	6596888 (83.44%)	6230939 (83.76%)	9275868 (81.97%)	10355298 (82.74%)
Metabolism	476570 (6.02%)	437987 (5.89%)	733393 (6.48%)	793275 (6.34%)
Unclassified	452153 (5.72%)	418420 (5.62%)	725631 (6.41%)	734914 (5.87%)
Environmental Information Processing	174422 (2.21%)	161246 (2.18%)	269884 (2.38%)	292322 (2.33%)
Genetic Information Processing	112890 (1.43%)	104151 (1.4%)	165044 (1.49%)	182698 (1.46%)
Cellular Processes	44943 (0.56%)	41927 (0.56%)	70619 (0.62%)	77593 (0.62%)
Human Diseases	31545 (0.39%)	29168 (0.39%)	50322 (0.44%)	51454 (0.42%)
Organismal Systems	16476 (0.21%)	15144 (0.20%)	25149 (0.22%)	27878 (0.22%)
No Hits	0	0	0	0

Table 13 – Tabulated data showing the number of matching reads to KEGG hierarchy pathway system from four datasets (Dataset 1 – MiSeq Nextera, Data 2 – MiSeq Nextera-XT, Dataset 3 – MiSeq NEXTFlex PCR Free and Dataset 4 – Ion-Torrent PGM). Please note the percentage of sequence reads matched to each pathway was calculated by dividing the individual reads from each pathway by the overall total number of reads from each dataset.

Summary of all metagenomics reads assigned to KEGG orthology

Sample	Sample ID	Read Length (bp)	Total Reads used for KEGG analysis	No of assigned read to KEGG hierarchy	Matching KEGG (%)
Illumina Platform	MiSeq Nextera	151	7905887	1308999	16.56%
	MiSeq Nextera-XT	151	7438982	1208043	16.24%
	MiSeq Nextflex PCR-free	251	11315910	2040042	18.02%
Life-Technologies PGM	Ion-Torrent PGM	400	12515432	2160134	17.26%

Table 14 - Summary statistics for the number of metagenomic sequences used for the KEGG analysis in this project. The percentage of matching KEGG was calculated from the number of reads assigned to all KEGG hierarchy divided by the total number of reads used in the KEGG analysis.

The number of sequencing reads from our datasets that could be assigned, were classified into seven default KEGG categories: metabolic pathways, environmental processing, genetic processing, cellular processing, organismal systems, human diseases and unassigned or unknown ambiguous sequences. A summary of the assigned reads used in this analysis are presented in Figure 46. As expected, the 'unclassified category' for all datasets, contained the highest number of sequencing reads due to many ambiguous low matching protein sequences. Genes involved in the 'metabolism' hierarchy such as the carbohydrate and energy metabolism, amino acids synthesis and lipid, and the glucose metabolism made up the next highest category, with more than 20% of the sequencing reads from all datasets (Nextera: 6.02%, Nextera-XT: 5.89%, PCR-free: 6.48% and Ion-Torrent: 6.34%) assigned to this category (Table 13) (Figure 46).

Genes involved in 'environmental information processing' accounted for more than 15% for all datasets (Nextera: 2.21%, Nextera-XT: 2.18%, PCR-free: 2.38% and Ion-Torrent: 2.33%) and this category contains the membrane transport signalling molecules which are important in many bacterial secretion systems (Figure 46). Next, both "cellular organismal systems" and "human diseases" categories which contain vital information for bacterial cellular activity such as "toxic-secretion" were represented by approximately 1% of the sequencing reads (Nextera: 0.95%, Nextera-XT: 0.95%, PCR-free: 1.06% and Ion-Torrent: 1.04%) assigned to the KEGG orthology system. These assignments are particularly important as they provide insight into bacterial properties especially concerning their virulence towards human hosts and other possible vertebrates (Table 13) (Figure 46). In addition, Table 14 shows the comparison between MiSeq and Ion-Torrent assignments together with their matching KEGG percentage scores. The results are comparable.

To further investigate the metabolic activities of the microbial community approximately 3,107,354 reads that had been assigned to the metabolism pathway were further divided into two networks, "carbohydrate" and "energy metabolism" (Figure 47). Further expansion of these two nodes revealed another seven available sub-classification groups shown in Figure 47 along with their respective assigned sequencing reads. These groups comprise a "glycolysis" subgroup, represented by 170,645 reads, a "citrate/TCA" cycle subgroup with 170,333 reads, a "fructose and mannose metabolism" subgroup with 67,152 reads, an "amino sugar and nucleotide sugar metabolism" subgroup with 113,412 reads, an "oxidative phosphorylation" subgroup with 204,071 reads, a "carbon fixation" subgroup with 200,584

reads and finally a "nitrogen metabolism" subgroup with 224,928 reads (Figure 47). Other metabolic activities were also present, but are not reported here, as analysis of the above seven groups provides sufficient data to make some statements about the metagenome datasets.

In respect of bacterial pathogenesis the sequences from 208,755 reads matching the KEGG "human disease" pathway node were further divided into six categories. These included a "*Vibrio cholerae* infection and pathogenic cycle" category with 21,405 reads, an "epithelial cell signalling category for *Helicobacter pylori* infection" with 12,908 reads, a "*Salmonella* infection" category with 8,796 reads, a "*Bordetella pertusis*" (whooping cough) category with 22,826 reads, a "Legionellosis" (Legionella spp) category with 38,136 reads and lastly a "Tuberculosis" group with 33,518 matching reads respectively (Figure 48). There were also other subgroups with very low levels of representation in our KEGG analysis. Despite these indications of potential health concern, most of the assigned sequencing reads in our datasets were too low (insufficient) to unambiguously determine pathogenic strains of disease and thus our study only hints at the importance of additional sequencing and analysis.

From the combined dataset of metagenomic community samples we used the KEGG analysis mapping tool in MEGAN5 software to analyse reads matching the Tricarboxylic Acid (TCA) and nitrogen cycles. This was done to obtain a deeper understanding of energy metabolism, nitrification and denitrification processes (Figures 49 and 50). Based on the KEGG analysis and TCA metabolic chart, we identified several enzymes (Table 15) exclusively present in our metagenomic datasets which are responsible for the utilization of adenosine triphosphate (ATP) (Figure 49). These enzymes were identified as phosphoenol pyruvate [EC: 4.1.1.32], 2-hydroxylethyl-THPP, [EC: 1.2.4.1], cis-aconitate [EC: 4.2.1.13], 3-carboxy-1-hydroxypropyl-THPP [EC: 1.2.4.2], lipoamide-E [EC: 1.8.1.4], fumarate [EC: 1.3.99.1] and s-malate [EC: 4.2.1.2] (Figure 49) (Table 15). These enzymes utilise many organic compounds (heterotrophic) in the production of energy, through processes such as gluconeogenesis, lipid metabolism and amino acid metabolism via hydrolysis. As for denitrification and nitrification processes, we assigned 224,928 sequencing reads to the nitrogen reduction and fixation map via the KEGG orthology (KO) module with MEGAN5 software (Figure 50). Similarly for the TCA metabolic chart, we identified several enzymes involved exclusively in nitrification and denitrification processes. These analyses identified enzymes important in the biosynthesis of carbamoyl phosphate synthetase [EC: 2.7.2.2],

5,10-Methylenetera hydrofolate [EC: 2.1.2.10], Urocanate [EC: 4.3.1.3], Orthophosphate [EC: 6.3.1.2], as well as 2- oxidase oxoglutarate [EC: 1.4.1.1.3], Nitrous oxide reductase [EC: 1.7.99.6 and 1.7.99.7] and Hydroxylamine oxidase [EC:1.7.1.4, 1.7.7.1 and 1.7.2.2] (Figure 50) and (Table 16). These enzymes are responsible for processing many amino acids such as arginine, proline, glycine, histidine, aspartate, glutamate as well as other nitrogenous compounds.

Vibrio cholerae infection and pathogen cycle

In addition, the analysis of our combined Illumina metagenomic DNA dataset indicated that approximately 21,405 (13.74%) reads out of 155,769 sequencing reads were assigned to the "*Vibrio cholerae* infection and pathogenic cycle" categories under the "infectious disease" KEGG pathway node (Figure 51 and 52). In the *Vibrio cholerae* pathogenesis map, the pathogenic cycle was divided into three phases of gene regulation: pre-exponential, stationary and post-exponential. In the pre-exponential phase, about 2,301 and 3,996 sequencing reads (Table 17) were assigned to *FlrA* [KO 10941] and *RpoN* [KO 3092] respectively, which are genes for bacteria chemotaxis and flagellar assembly (Figure 51). Next for the stationary phase, sequencing reads were allocated to *RpoS* [KO 3087; 2167 reads], *CRP* [KO 1587, 1587 reads] and *AC* [KO 5825, 5825 reads] (Table 17). These are classified as class III genes belonging to actin assembly-inducing proteins responsible for bacteria motility such as flagella assembly, construction of basal body and motor components (flagellins) (Figure 51). Lastly for the post-exponential phase (which is which concerns genes for toxicity, approximately 1,167 sequencing reads (Table 17) were assigned to the *ToxT* [KO 10923] (Figure 51) gene which is responsible for regulation of pH-gradients.

Preferentially selected TCA cycle enzymes from the microbial community found in our metagenomics datasets along with their essentiality/primary function

KEGG Orthology (KO) number	Enzyme nomenclature (EC) number	No of assigned sequencing reads	Enzyme Identification (ID)	Essentiality
KO 1596	4.1.1.32	2588	*phosphoenol pyruvate*	Gluconeogenesis
KO 1610	4.1.4.49	4986	*phosphoenol pyruvate*	Gluconeogenesis
KO 0161 – 0163	1.2.4.1	21457	*2-hydroxylethyl-THPP*	Fatty acid/lipid metabolism
KO 1681 – 1682	4.2.1.13	24765	*cis-aconitate*	Amino acid metabolism
KO 0164	1.2.4.2	15024	*3-carboxy-1-hydroxypropyl-THPP*	Amino acid metabolism
KO 0382	1.8.1.4	11133	*lipoamide-E*	Amino acid metabolism
KO 0239 – 0247	1.3.99.1	19896	*Fumarate*	Amino acid metabolism
KO 1676 – 1679	4.2.1.2	15260	*s-malate*	Amino acid metabolism

Table 15 – Enzymes found in our metagenomic datasets from the TCA pathway, along with number of sequencing reads assigned to KO and EC numbers

Essential enzyme associated with KEGG Nitrogen Metabolism hierarchy

KEGG Orthology (KO) number	Enzyme nomenclature (EC) number	No of assigned sequencing reads	Enzyme Identification (ID)	Essentiality
KO 0926	2.7.2.2	1700	*carbamoyl phosphate*	*Arginine* and *proline* metabolism
KO 0605	2.1.2.10	3806	*5,10-Methylenetera hydrofolate*	*Glycine* metabolism
KO 1745	4.3.1.3	9618	*Urocanate*	*Histidine* metabolism
KO 1915	6.3.1.2	24282	*Orthophosphate*	*Alanine, aspartate and glutamate* metabolism
KO 0264-266	1.4.1.1.3	35101	*2- oxidase oxoglutarate*	*Alanine, aspartate and glutamate* metabolism
KO 4561, 2305, 2448, 2164, 4747 and 4748	1.7.99.6 1.7.99.7	11305	*Nitrous oxide reductase*	Nitrogen metabolism
KO 0362 and 0363	1.7.1.4 1.7.7.1 1.7.2.2	8169	*Hydroxylamine oxidase*	Ammonia metabolism

Table 16 – Metabolic enzymes responsible for nitrogen metabolism found in our metagenomic sample with KEGG orthology and enzyme nomenclature (EC) numbers and their essentiality.

Most represented KEGG functional assignment hierarchies

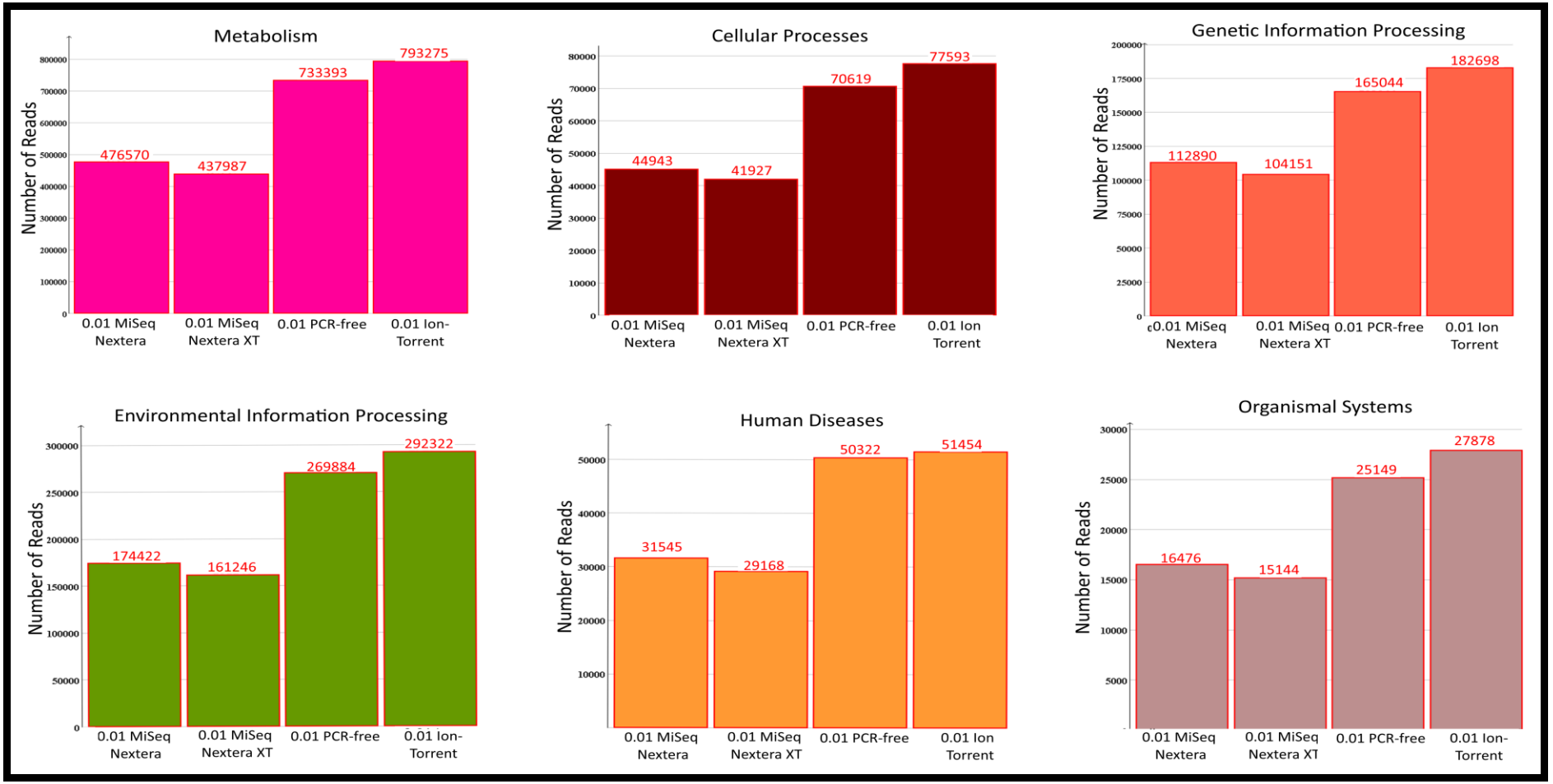


Figure 46 – KEGG analysis for Tamaki River showing main classifications of functional content for metagenome datasets. This analysis involved assigning sequence reads to KEGG orthology categories - metabolism, cellular processes, genetic information, environmental processing, risk of human disease and organismal system. There was no normalisation of the sequences for this data analyses.

KEGG 'metabolism' hierarchy pathway

Figure 47 – KEGG analysis for the "metabolism" pathway indicating "carbohydrate and energy metabolism" categories and subcategories. The relative number of reads assigned to different functional nodes is shown. The figures in red indicate the number of sequencing reads assigned to the respective functional content network by the KEGG orthology system**.** Here we have chosen six important categories for the functional analysis: Gluconeogenesis, the Citrate cycle, Fructose and Mannose metabolism, amino and nucleotide sugars metabolism, oxidative phosphorylation and finally the Nitrogen metabolism.

KEGG 'human disease' hierarchy pathway

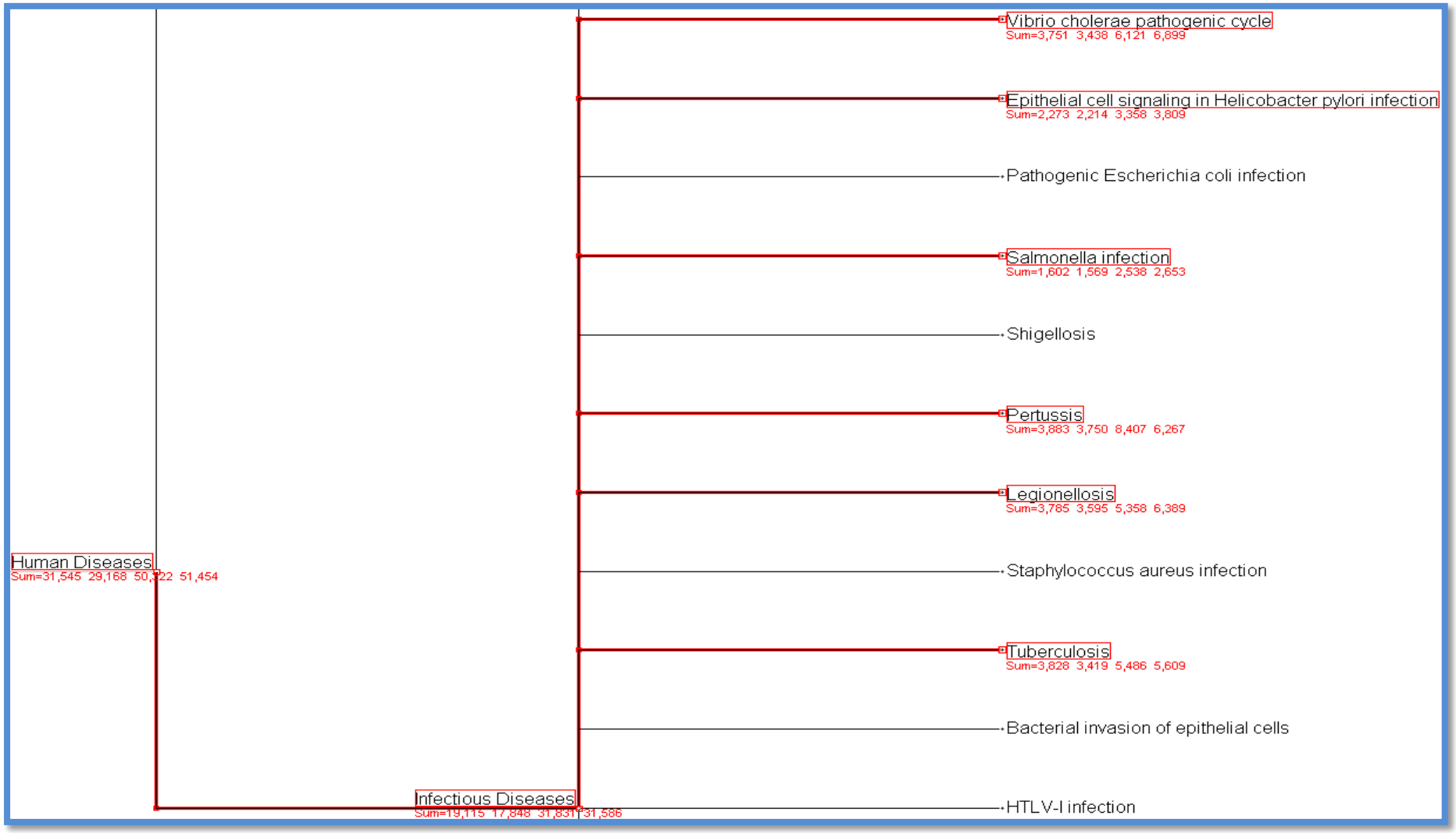


Figure 48 – KEGG analysis on "Human Disease" functional hierarchy indicating potential associations of infectious disease from our metagenomics datasets. Functional nodes highlighted in red show six potential infectious diseases (*Vibrio cholera, Helicobacter pylori, Salmonella, Bordetella pertussis, Legionella and Mycobacterium tuberculosis*) of interest in our project together with the number of sequencing reads assigned to it. Most of the assigned sequencing reads are very low due to lower coverage and average quality sequences.

Vibrio cholerae pathogenicity factors

For pathogenicity factors such as the interaction between *Vibrio cholerae* and its human host we only have about 1,466 matched reads from all datasets (not shown) assigned to KEGG due to lack of sequence reads. However the *Vibrio cholera* infection cycle pathway has a higher number of assigned reads (11,654 reads from all datasets) (Figure 52). The pathway information is important as it involves numerous genes and enzymes responsible for the invasion of host cells (Figure 52). Most of these genes and enzymes inhibit and disrupt the ion transport system such as calcium uptake, water and electrolyte secretion systems, colonization and activation of endocytosis which causes actin polymerization (Figure 51 and 52). We identified some of the genes and enzymes in our datasets important for bacterial infection cycles. These were: V-type H^+ subunit-A (hemolysin) [KO 00190], Protein transport SEC61 subunit-alpha complex [KO10956], Protein Kinase-A (PKA) [KO 4345], RTX toxin-A [KO 10953], actin-beta gamma 1 (G-actin) [KO 5692] and Vibrio Lysine [KO 8604] (Figure 51 and 52) (Table 18). However, most of our assigned reads to the KEGG orthology were very weak and there are too few genes identified to assess the overall pathogenicity of *Vibrio* in the Tamaki River sample.

Summary of matched genes found in *V. cholerae* pathogenesis pathway

KEGG Orthology (KO) number	Genes	No of assigned sequencing reads	Phase	Functionality
KO 10941	*FLrA*	2301	Pre-exponential	Class I genes for chemotaxis protein
KO 3092	*RpoN*	3996	Pre-exponential	Class III genes for basal body and motor components
KO 2405	*FLiA*	2103	Pre-exponential	Class IV genes for motor components
KO 3087	*RpoS*	2167	Stationary	Attachment of bacterial cells to host epithelial cells for invasion
KO 10914	*CPP*	1587	Stationary	Carbon synthesis for cyclic AMP process
KO 5831	*AC*	5825	Stationary	Carbon source and temperature regulator
KO 10923	*ToxT*	1167	Post-exponential	Toxicity expression, type II secretion system

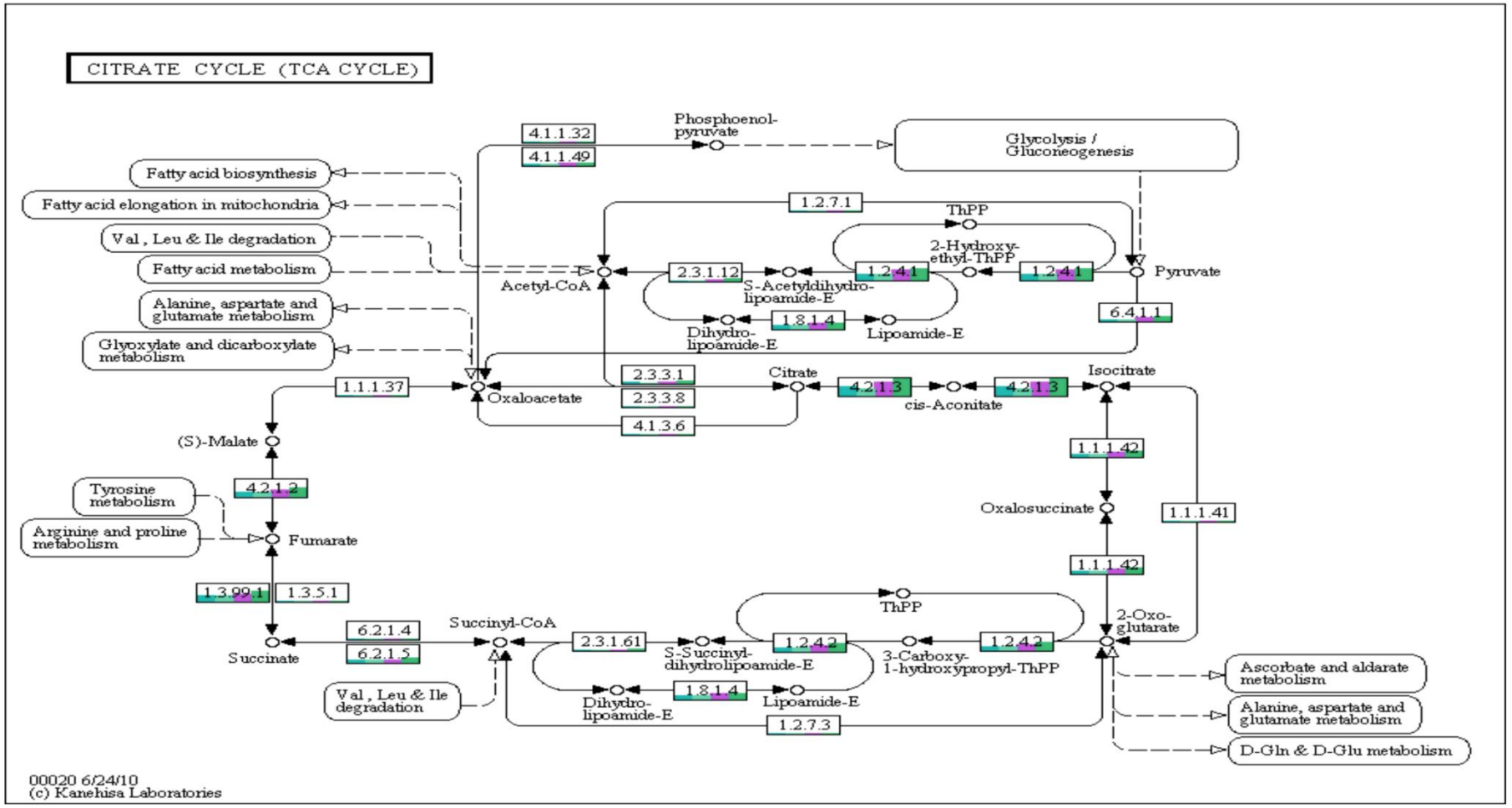
Table 17 – Genes associated with *V. cholerae* pathogenesis and its functionality in our metagenomics datasets.

Summary of matched genes and enzymes found in *V. cholerae* infection pathway

KEGG Orthology (KO) number	Genes and Enzymes	No of assigned sequencing reads	Functionality
KO 00190	*V-type H^+ transporting ATPeV1A subunit (hemolysin)*	184	Class I genes for chemotaxis protein
KO 10956	*Sec61 (protein transport subunit)*	64	Class III genes for basal body and motor components
KO 4345	*PKA (protein kinase A)*	42	Class IV genes for motor components
KO 10953	*rtxA toxin*	277	Attachment of bacterial cells to host epithelial cells for invasion
KO 5692	*G-actin (actin-beta gamma 1)*	45	Carbon synthesis for cyclic AMP process
KO 8604	*Vibrio lysine*	59	Carbon source and temperature regulator

Table 18 – Genes and enzymes from our metagenomics dataset linked to the *V. cholerae* infection pathway.

KEGG Citrate Acid Cycle Metabolic pathway chart mapped from Tamaki River reads

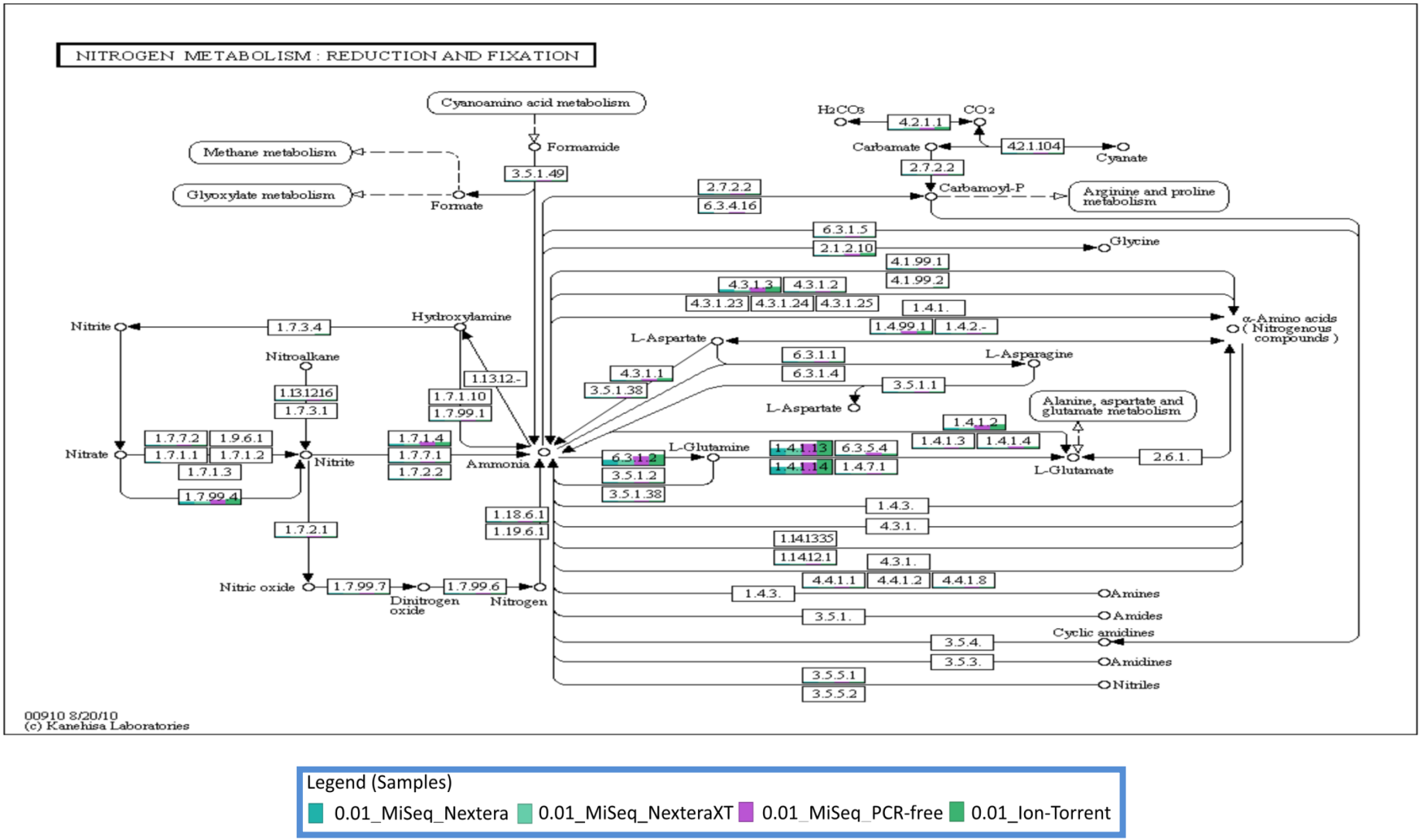


Legend (Samples)

0.01_MiSeq_Nextera 0.01_MiSeq_NexteraXT 0.01_MiSeq_PCR-free 0.01_Ion-Torrent

Figure 49 – Tricarboxylic Acic Cycle (TCA) or also known as Krebs cycle is an important metabolic pathway for generation of energy in many bacterial species. The figure shows genes mapped to the TCA pathway from our metagenomics reads.

KEGG Nitrogen Metabolism pathway chart mapped from Tamaki River reads

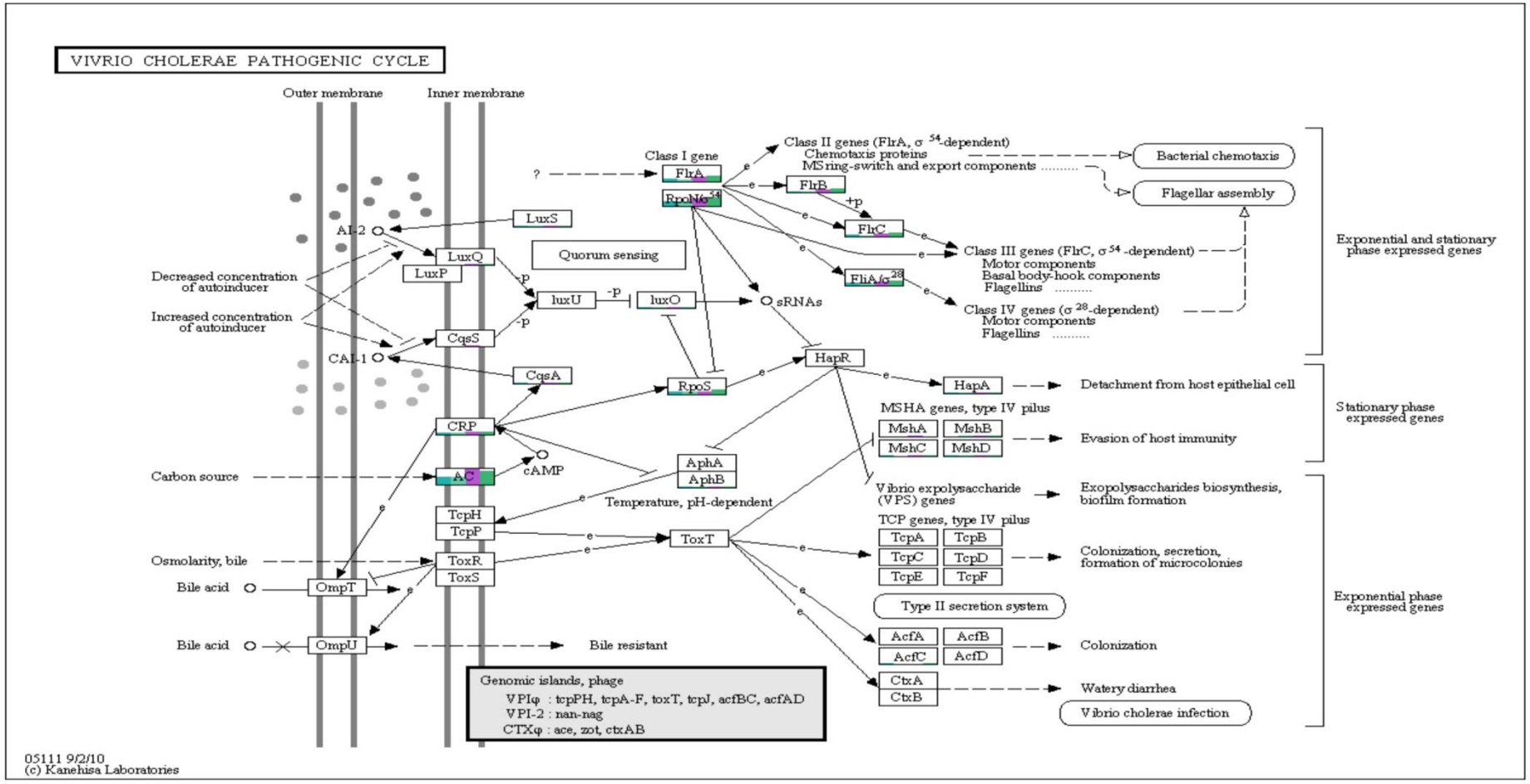


Legend (Samples)

0.01_MiSeq_Nextera 0.01_MiSeq_NexteraXT 0.01_MiSeq_PCR-free 0.01_Ion-Torrent

Figure 50 – The nitrogen metabolism cycle is used by many bacteria for processing organic and inorganic nitrogen compounds for ammonification, mineralisation, nitrification and denitrification processes. Reads mapping to key pathway steps have been indicated.

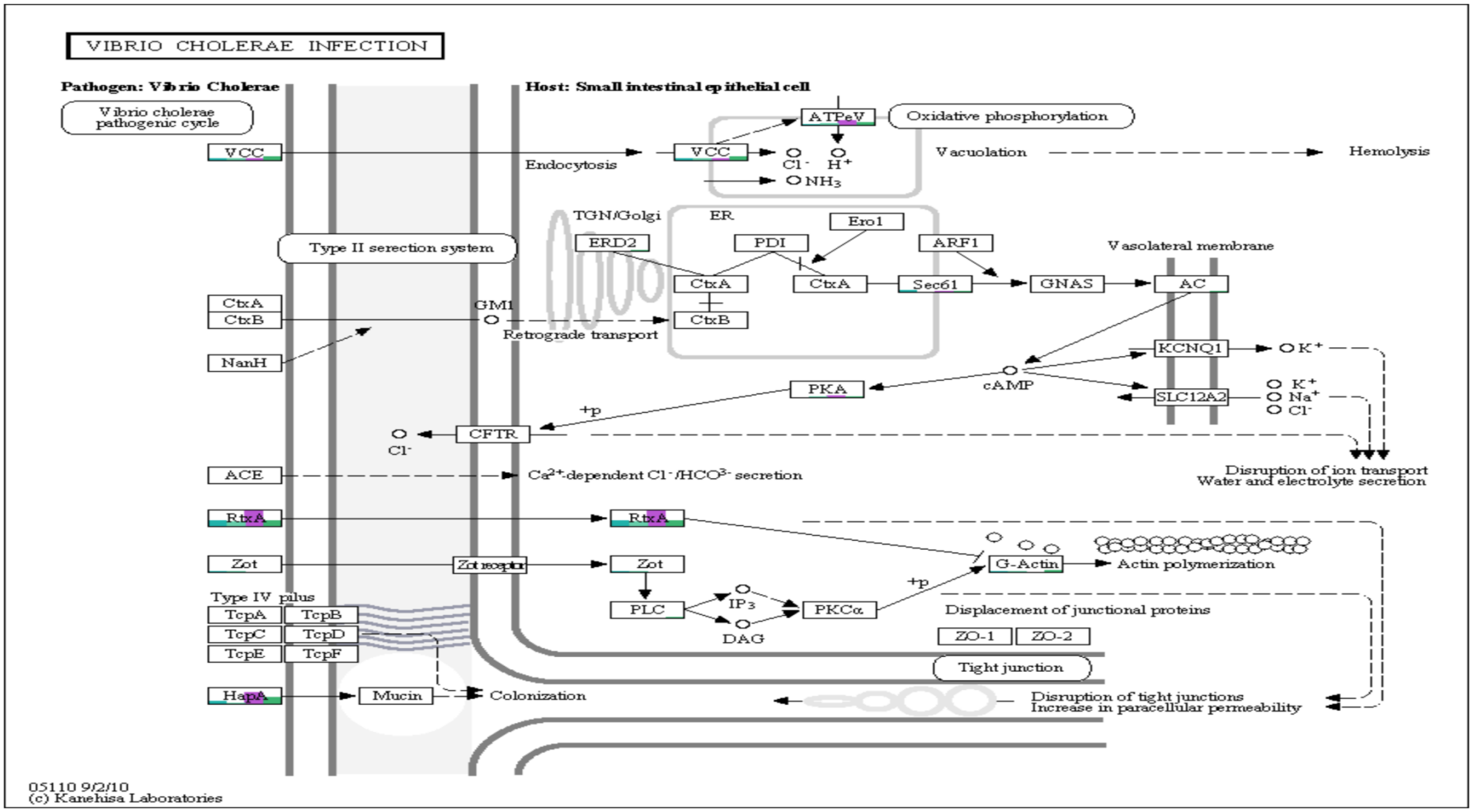
KEGG *V.cholerae* pathogenic cycle pathway from Tamaki River reads



Legend (Samples)
0.01_MiSeq_Nextera 0.01_MiSeq_NexteraXT 0.01_MiSeq_PCR-free 0.01_Ion-Torrent

Figure 51 – Pathogenesis-associated colonization of *V.cholerae* cycle. This cycle shows the dual life cycle of *V.cholerae* in the aquatic environment and in the host during virulent phase when colonizing the human small intestine.

KEGG *V.cholerae* infection pathway mapped from Tamaki River reads



Legend (Samples)

0.01_MiSeq_Nextera 0.01_MiSeq_NexteraXT 0.01_MiSeq_PCR-free 0.01_Ion-Torrent

Figure 52 – *V.cholerae* infection pathway (in human) indicating the steps required for virulence i.e. secretion of Cholera toxin (CTX). The highlighted area is where our metagenomics reads have been mapped.

Functional microbial attributes

Finally, the combined reads from different NGS sequencing protocols were assigned to 'functional microbial attributes' in MEGAN5. This assignment made use of the NCBI prokaryotic attribute table (Huson et al., 2009). Sequencing reads were assigned and grouped into fifteen prokaryotes attribute categories (including metabolism, virulence factor, pathogenic properties and habitat). Figure 53 shows the potential relationship between the functional groups to which reads were assigned. The microbial attribute analysis indicated that most of our taxa belonged to an aquatic ecosystem habitat, and thrived at optimal temperatures between 28ºC to 30ºC, and comprised opportunistic bacteria (Figure 53).

KEGG co-occurrence microbial relationships plot in Tamaki River

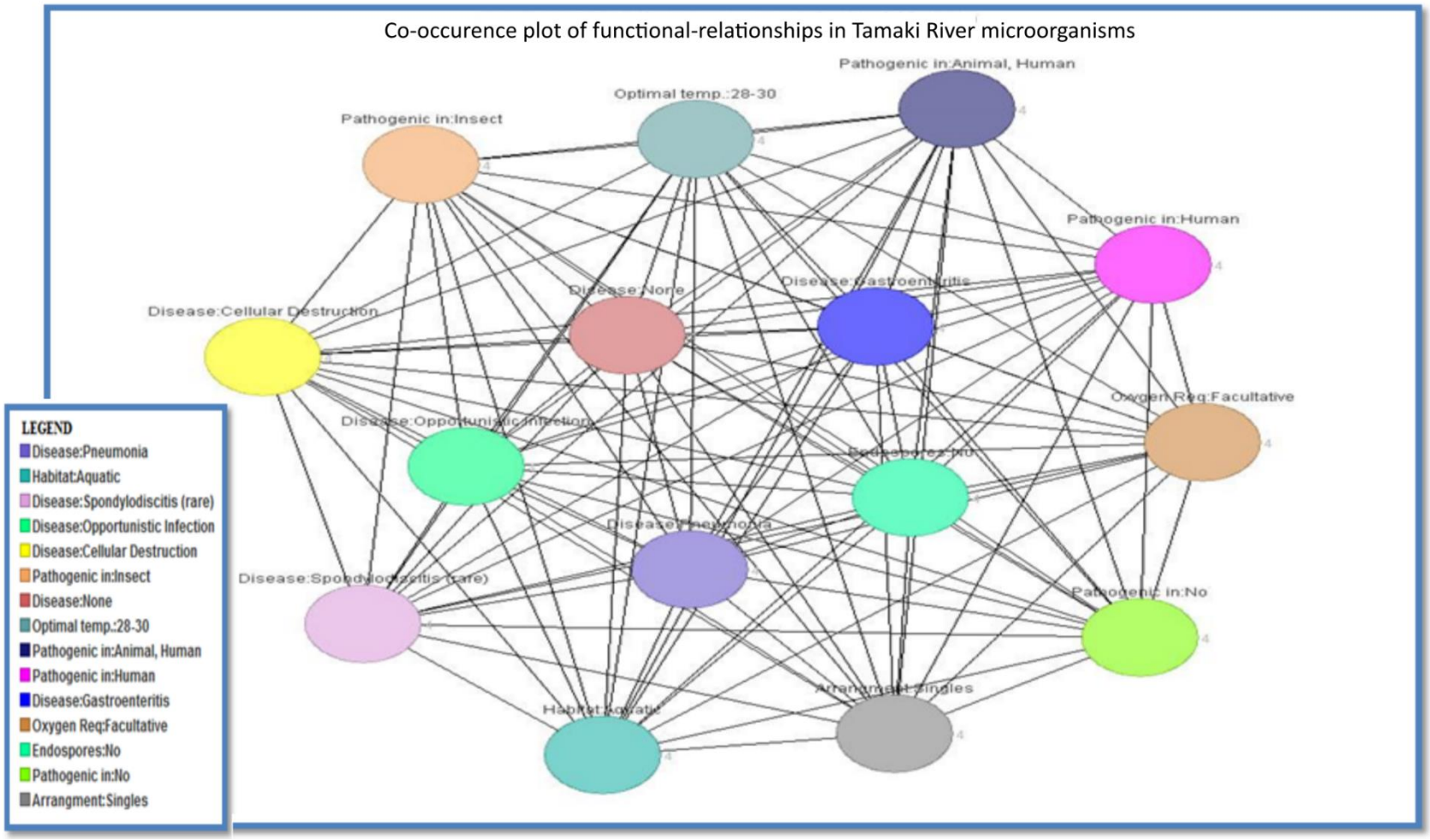


Figure 53 – Microbial attributes co-occurrence chart plotted from KEGG analysis classification based on reads from MiSeq Nextera, MiSeq Nextera-XT, MiSeq NEXTFlex PCR Free and Ion-Torrent PGM sequencing protocols. The chart indicates common functional gene group relationships in the Tamaki river microorganisms.

4 Discussion

4.1 Sampling and filtration strategy

For metagenomics analysis, where we seek to survey and understand complex microbial communities in different habitats, the sampling strategy is an important factor for consideration. Factor such as type, size, scale and timing of the collected samples can answer key questions such as: (1) what (microorganisms) are present in the environment, (2) what are they doing there, and (3) how do they react to a certain environmental changes? (Cantarel et al., 2011; Press, 2007). Technical issues such as sample collection, extraction, library preparation protocols, sequencing method and computational annotation may also influence the downstream analysis of the metagenomics data (Raes et al., 2007). Sample collection and processing are the preliminary steps in any metagenomics project and collected samples must be monitored and tightly controlled to prevent contamination (Thomas et al., 2012). The collection site, or habitat selection, was important for this project and in both cases we chose a freshwater habitat (a river) from a farming region of New Zealand. Freshwater is an important resource for agriculture especially when it is used for irrigation to provide water for open fields growing vegetables, fruits, grains and as a water source for domestic animals. The Tamaki River location for sampling had previously been identified by Ministry of Primary Industry (MPI) as within a "high-risk" zone for *E. coli*, and also a location where there was also a high chance of encountering gastroenteritis disease-causing organisms such as *Campylobacter jejuni*, *Giardia lamblia* and *Cryptosporidium parvum*.

To ensure proper collection and sampling of the microbial biodiversity within this environmental ecosystem, a robust and sustainable sampling method was needed. In our study, the water samples were treated with the utmost stringency to ensure no contamination was introduced during sampling and DNA extraction protocols. Measures included the single use of water collection and filtration tools, sterilising (by autoclaving) of water storage bottles and of the filtration apparatus between collections, storage of the samples at 4°C and filtration within 24 hours. In the preliminary stage of our project, we investigated water sampling using a "stomacher". However even taking care and using the strictest protocols, we still experienced minor cross-contamination with our first collection of samples processed using the stomacher (Agency., 2001). This was indicated by an unexpectedly high proportion of '*Psedomonadales*' from the *Proteobacteria* family identified in metagenome analyses of

the stomacher samples. This problem occurred due to an inefficient cleaning process and handling of the stomacher filtration apparatus. For this particular reason we abandoned the use of the stomacher and recollected the water samples as one litre 'grab' water samples. Figure 54 provides an overview of the workflow used in this project from water collection to data analysis.

Work flow used for water screening

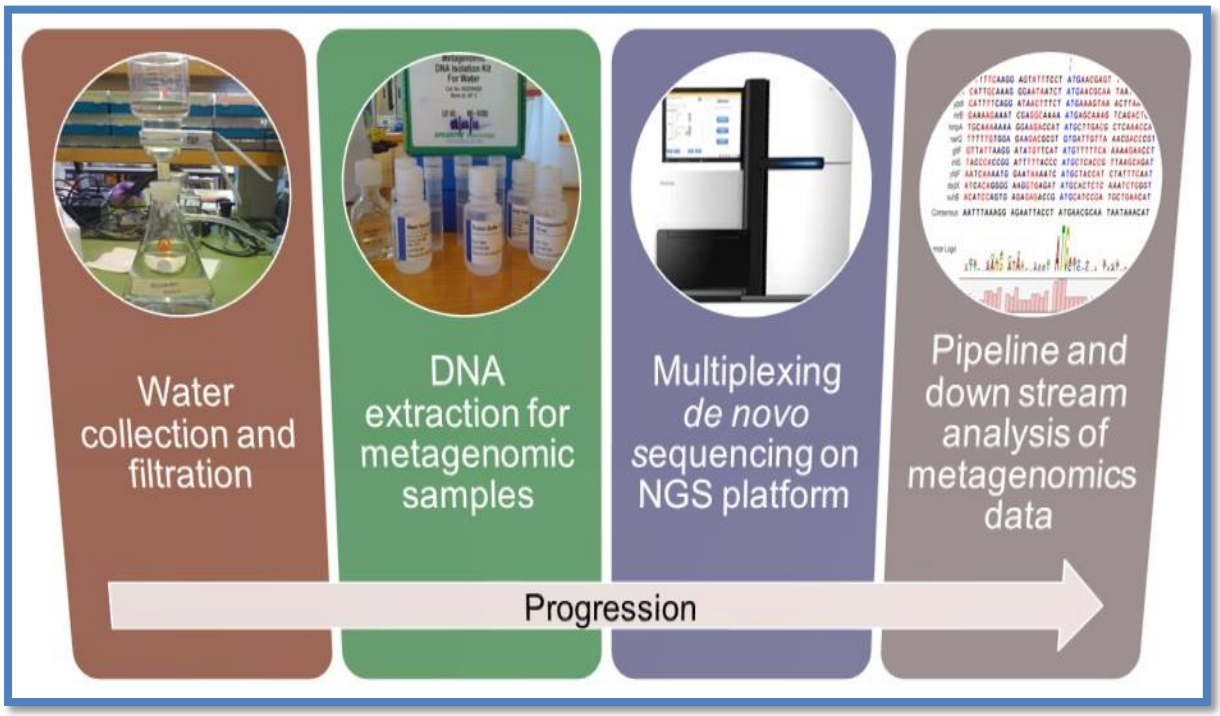


Figure 54 – The workflow above shows the main steps followed in the current project. Different library preparation protocols were used for NGS sequencing. All were able to detect a wide range of microbial species.

4.2 Optimization of NGS library preparation workflow.

4.2.1 Overcoming poor DNA yields from low biomass samples.

Although metagenomics is potentially a powerful methodology, it can be limited by the requirement for sufficient sample quality and quantity, cross-contamination of filtrates, sample-bias and the time needed for database matching and downstream sequence analysis (Amorim et al., 2008; Rachel et al., 2014).

Unlike marine ecosystems, our project involved freshwater habitats where microbial diversity are much more complex and diverse. A mixture of soil, waste and living organisms provides many nutrients and serves as a reservoir for many different types of microbial communities (Bertrand et al., 2005; Eichler et al., 2006; Tringe et al., 2005). Water collection through different pore-size filters and tangential filtration are commonly used methods for microbes for aquatic metagenomics research (Cottrell, Matthew T. et al., 2005; Djikeng et al., 2009; Rusch et al., 2007; Venter, J. Craig et al., 2004).

At an early stage of this project we utilized a filtration technique known as tangential filtration or cross-flow filtration. In this type of filtration, the water is pressurized and forced through a specific permeable membrane bed to capture the solid or protein impurities and purify the filtrate (Djikeng et al., 2009). This method requires that the filters are further extracted by a machine (a stomacher). After concentration and purification by the stomacher instrument, the water samples are filtered. Filtration involves a pre-filtration step using a cheesecloth or larger porosity filters (20 µm) for removal of larger particles and unwanted waste. A second step then uses smaller pore-sized filters (0.1 to 5 µm) to capture the microbes of interest. Although use of the stomacher ensures a much greater volume of water can be filtered, we encountered problems with contamination in the earlier collected samples. Part of this problem arises due to the sensitivity of Next Generation Sequencing (NGS). That is, cleaning methods that may have been sufficient for previous microbiological testing are no longer stringent enough when NGS protocols are employed.

High quality molecular weight (hmwt) DNA (OD260/280 of 1.8 to 2.0, total >1 µg) is crucial for library construction as it provides better fragmentation, ligation and a more even distribution of sequenced reads (Kakirde et al., 2010; Krsek et al., 1999; Lemarchand et al., 2005; Wommack et al., 2008). For library optimization, our DNA extraction protocol was divided into three steps; the washing of the filter papers, the enzymatic reactions (cell lysis and extraction) and the purification of the extracted genomic DNA. To extract the genomic DNA, the filters were agitated vigorously to wash off any biofilm, or solid sediment, that sticks to the surface of the filter paper. To aid this process addition of a washing agent such as 0.1% Polysorbate 20 (Tween-20) detergent is incorporated into the washing buffer (Shen et al., 2011) (Linke, 2009). The addition of detergents helps to release any existing hydrophobic material, protein and macromolecules during the filtration wash process.

The processes of cell lysis and extraction of nucleic acids are broken down into two parts: (1) cell lysis and the disruption of the cellular and nuclear membranes of the microbes to release the nucleic acids and, (2) separation of the DNA from the cell debris and other materials such as soil particles (Robe et al., 2003). Cell lysis and disruption of cell membranes can be achieved by utilizing several techniques such as (1) physical disruption, where physical force such as freezing-thawing, mortar grinding or bead-bashing are used (More et al., 1994; Tsai et al., 1991), (2) chemical lysis using detergent sodium dodecyl sulphate (SDS), Chelex-beads, CTAB (cetyltrimethyl-ammonium bromide) and/or PVPP (polyvinylpolypyrrolidone) which can require heat-treatment and the addition of EDTA as a chelating agent (Herron et al., 1990; Jacobsen et al., 1992; Nannipieri et al., 2006) and lastly (3) enzymatic reactions such as addition of lysozyme and proteinase K for cell lysis before the DNA is purified (Maarit Niemi et al., 2001; Tebbe et al., 1993). Initially for our DNA isolation method, the filter paper was first washed with physical force (bead-bashing) in a Magna-Lyser instrument (Roche). Unfortunately the genomic DNA isolation for this method was not promising as we only obtained a low amount of hmwt DNA which was not sufficient for our research project. We suspect at this stage the genomic DNA was over sheared due to the action of the 'hydrodynamic' forces on the DNA. In future, it would be interesting to run a series of random and non-random DNA fragmentations to investigate whether we achieve a uniformity in genomic fragment size without producing any biases.

Next we decided to try a commercial kit known as the "metagenomic DNA isolation kit" from Epicentre (Illumina) for chemical lysis and DNA extraction. This protocol is fast (Murray, 2008) and requires no additional special reagents for mechanical cell disruption and isolation, thus using it simplified the overall process by reducing the amount of time required for the extraction process. Here the separation of the DNA from the cell debris and other materials is achieved by binding the DNA to a silica-based type column with high salt buffer, prior to washing the column with salt/EtOH buffer and eluting the DNA in a low salt buffer (usually 10mM Tris or water). Column technology utilises the property of DNA to bind to silica under high salt conditions. The column serves to clean and concentrate the DNA. While the DNA is bound to the column, most of the contaminants can be washed away. The binding reaction is reversed at low salt concentration.

To maximise the DNA recovery yield, we also used multiple filters for each one litre grab sample. We found this gave us a 3-fold increase in the total amount of DNA that could be

recovered. Although some of the library preparation methods only required a small amount of genomic DNA template as starting material for next-generation sequencing, nonetheless, a good quantity and quality of DNA is always desirable as the additional extracted DNA can serve as for additional analyses. Sufficient template for library construction contributes towards the overall success of the high-throughput reads obtained from the sequencing run. Our optimization results show that there was a strong correlation between the concentration of the starting material and the shearing efficiency, where a higher amount of starting material always sheared better and gave a narrower band of interest, compared to a lower amount of genomic DNA. This finding correlates with the findings of other researchers. For example, a paper published by Aird and colleagues has shown that increasing the amount of DNA material for NGS library construction to about twice the recommended specification on the Illumina TruSeq DNA protocol can greatly reduce the selection biases associated with PCR enrichment from the ligated libraries (Aird et al., 2011).

4.2.2 Issues with the next-generation sequencing library preparation protocols

Whole metagenome or shotgun metagenomic sequencing is now the most widely used method for investigating biodiversity in an ecosystem. In this part of the study, we examined and investigated whether next-generation sequencing library preparation impacted on environmental data. Factors such as DNA yield and purity, robustness of the preparation kit, optimization and read bias were explored.

In next-generation sequencing there are several ways to construct a library and choices regarding which preparation kit to use are most commonly made available based on the amount and purity of the DNA material. There are numerous commercially available sequencing methods and platforms for metagenomics sequencing and each of them have their own designated requirements.

Here we investigated four library preparation methods on data quality for comparative studies based on library complexity. These were (1) Nextera DNA and (2) Nextera-XT DNA, (3) NEXTFlex PCR-free and lastly the (4) Ion-Torrent Xpress 400 bp kit. There are several external factors that can significantly affect the process for NGS library construction and each of them plays a crucial role in producing good quality data with a high yield (output). The construction of the library can be greatly influenced by a variety of conditions known as “batch effects”. This refers to the laboratory practices and conditions such as temperature,

humidity, workflows, variation in batches of reagents and concentration differences which may occur during the library construction process. These influences include the preparation time, poor coverage, complexity, cost, reagent robustness, and consistency and are only a few examples of such pitfalls that will be further discussed along with some potential solutions.

The preparation time from each commercial kit from different companies can vary from several hours to days to generate NGS samples ready for sequencing. Poor fragmentation accounts for some poor quality reads and can cause biases in obtaining metagenomics reads (Poptsova et al., 2014). Improper shearing due to insufficient or overtime fragmentation can lead to cutting at the "preferred" positions on the genomic DNA resulting in generating inaccurate lengths of fragments with non-random ends (Poptsova et al., 2014). This error can be a problem for downstream processes such as end-repair, A-tailing and adapter ligation steps, where an uneven distribution of fragmented ends can lead to low library complexity and sampling biases towards a specific "preferred" sequences. For example, we needed to modify the incubation time of 5 minutes documented for Nextera and Nextera-XT preparation protocols to about 12 to 15 minutes depending on the genomic DNA concentration, due to the poor fragmentation process. The enzymatic reaction was very unpredictable and on several occasions our generated "end-product" of NGS libraries was significantly larger than 2 kb which can be an issue for clustering during an Illumina sequencing run. A large genomic product will tend to over-cluster causing uneven light intensity during the sequencing run which can cause problems for the base-calling algorithm. Besides such poor image resolution tends to lead to uneven base-calling which lead to improper matric/phasing issue and poor reads. Thus it is very important to ensure the final library fragments are within an proper size and as in our project a 550 bp fragment is the most ideal and is sufficient to generate an optimal cluster density. In addition, to further validate our fragment size we ran the NGS library on the Bioanalyzer instrument.

Another issue we encountered during sample preparation is the reagent robustness. Since different manufacturers have their own proprietary reagents and quality check (QC) systems, certain biases and discrepancies might be present in our libraries. For example in our Illumina Nextera-XT preparation protocol, we evaluated the performance of the Nextera-XT library preparation kit based on a genomic DNA concentration range from 0.8 to 5.0 ng/µl as the starting material that was fragmented for 5 minutes. We observed some discrepancy in the size distribution of the fragmented genomic DNA and it varied accordingly to different

ranges of starting material (gDNA) concentration. For validation of gDNA concentration with approximately 1 ng/µl as recommended by the protocol or similar, we observed most of our fragmented products were sheared to a correct size of between 750 to 1,150 bp; meanwhile a higher concentration of gDNA produced a fragment size of more than 1,500 bp and above which is too large for efficient clustering on the MiSeq instrument. Because of this, for every single Nextera-XT sample preparation we thoroughly checked the concentration of the gDNA starting material and diluted the sample to 1ng/µl. Here, we assume a DNA concentration closer to 1ng and a larger insert-size of fragmented library has the correct molarity and size accessible to the enzyme 'Tn5-transposase' to function properly compared to the larger fragmented insert-size library of more than 1.5kb. Thus for best practice, it is important to ensure that the DNA concentration and purity is optimized prior to the fragmentation process (enzyme based) to improve the robustness of NGS library construction.

Although the traditional method of size-selection based on agarose gel electrophoresis is effective, it can be time consuming, costly and produces lower yields of DNA. In our NGS library preparation, we utilized an alternative size-selection protocol known as the SPRI beads clean-up method which omits the gel-extraction method by adding different ratio concentration of PEG/NaCl SPRI beads to DNA during size-selection step. Here, we used a ratio of approximately 0.6X of Ampure XP SPRI beads to 20 µl of reaction mix for size-selection of a 550bp insert. During this step, adapter dimers and the T-overhanging bases are removed as are larger and smaller size inserts. It is important that we removed these unwanted size fragments as they do not cluster-well and will interfere with the cluster density and read quality (Derek Campbell, Illumina, personal communication, 2012). A reduction in fragment size range improves the clustering density and sequence quality because clusters of uniform diameter are easier to detect (Derek Campbell, Illumina, personal communication, 2012). However we also learned that an incorrect ratio of PEG/NaCl beads added to DNA can surprisingly lead to minor biases such as 1) uneven size distribution, 2) high frequency of concatemers and chimaeric products and 3) unequal distribution of primer/adapter dimers (Derek Campbell, Illumina, personal communication, 2012). For example we encountered a small proportion of our sequencing reads (Nextera and Nextera-XT) do not pass the MiSeq QC filter (Q-score less than 10) due to presence of longer DNA fragments (>1.2kb) which led to poor clustering and base-calling efficiency. Further evaluation from the Bioanalyzer results

confirmed the existence of longer fragments which was likely arose from a poor fragmentation step.

Besides, we also discovered a minority of chimaeric products and concatemers in our NGS library suspected to be due to the earlier mentioned inefficiencies in the fragmentation step that are known to cause an uneven distribution of short- and long-inserts. Larger DNA inserts tend to join together during the end-repair process which may not be properly optimized for A-tailing thus may not properly ligate to the T-tailed adapters (Lodes, 2016). Adapter-ligation is a critical process in library preparation as it depends on the ligation efficiency of the adapter binding to the targeted NGS fragment. A failure in the adapter-ligation process can lead to confounded biases such as NGS libraries being contaminated with many chimaeras and concatemer products which complicate the computational downstream analysis. Poorly ligated libraries were also contaminated with primer-dimer and adapter sequences as indicated by the FastQC and SolexaQA data assessment. These contaminated reads can cause an uneven distribution of nucleotide bases such as an excessive GC- or AT-base composition which indicates uneven coverage and poor data quality (Lodes, 2016). The presence of such chimaeric reads if present in high number can mislead taxon diversity and abundance estimates (Ross et al., 2013). To reduce the potential occurrence of chimaeric reads with the NEXTFlex PCR-free protocol, we utilized a SPRI clean-up method with this protocol (Ross et al., 2013).

In summary, data quality did not differ significantly for the different Illumina protocols. That is, our PCR-free data showed many similarities to data produced from both Nextera and Nextera-XT protocols with approximately 3.8 million reads on a 2 x 250bp PE run. Besides we also observed similarity and consistency in the microbial diversity population between the taxon analysis of different Illumina protocols where bacteria species of *Pseudomonas fluorescens, Yersinia enterocolitica, Pseudomonas putida, Yersinia Pestis* and *Escherichia coli* were present across our library preparation methods. Such reproducibility in our taxon profiles indicated that the library preparation kits we used had only a minor effect on sample diversity but further investigation is still needed to differentiate between the sample type and the material abundance with PCR biases, as over-interpreting datasets can cause conflicting results, thus hindering the true microbial population. With respect to some of the chemistry differences between different NGS library preparation kits, abandoning PCR-enrichment (as in the NEXTFlex PCR-free protocol), increasing the denaturation temperature along with

addition of a denaturing agent such as DMSO or Betaine (as in the Ion-Torrent protocol) and using a high-fidelity enzyme or using primers with DNA melting agents to prevent poor sequencing coverage (as in Nextera and Nextera-XT) had little impact on the results we obtained. Further comparison of the taxonomic profiles for Nextera, Nextera-XT, PCR-free protocols revealed no significant differences in terms of sequencing quality or GC content. Although we expect the PCR-free protocol to reduce GC biases, the PCR vs. non-PCR datasets showed little difference in GC content that could be attributed to ligation bias. Other factors such as differences in workflow complexity (e.g., different types of reagents, methods and instruments) had apparently little impact on the results. These findings suggest that for evaluation and identification of bacterial diversity we do not need to normalize the sequencing protocol and platform we used, however in the author's personal opinion he suggested that a standardization of sequencing kits, platforms and practices should be adopted for improving the accuracy of metagenomics taxonomy assignment.

4.3 Performance comparison of Illumina MiSeq and Ion-Torrent sequencers

Here we will be comparing and discuss the data quality produced by MiSeq (2 x 150bp and 2 x 250bp protocols) and Ion-Torrent PGM (1 x 400bp protocol) along with the quantity of sequencing data. We also considered read length differences in data produced from each platform and the base-calling quality (including the occurrence of homopolymers).

Workflow

The Ion-Torrent has a much simpler/easier workflow compared to the MiSeq wherein a metagenomics library can be prepared in about 6 hours (using emulsion PCR). It also has a faster turnaround time for full data QC without much physical manipulation and this is the main selling point for the Ion-Torrent platform. This is possible as the Ion-Torrent utilizes technology based on chemical pH changes compared to MiSeq which uses the sequencing-by-synthesis method. Sequencing read length and rate of sequencing we obtained from the Ion-Torrent run was much faster with significantly longer reads with an average of read length of 400bp in under <4 hours whereas compared to Illumina MiSeq with an average read length of approximate 251bp generated in less than 24 hours per run. However a caveat with the Ion-Torrent data is that we had higher errors compared to the paired-end data from the

MiSeq as analyses could not discriminate the homopolymer variants efficiently. We observed that sequences in repetitive regions were poorly called as the Ion-Torrent has a difficulty in estimating the correct stoichiometric ratio/area during detection of H+ ion. This does not happen with the MiSeq sequencer as the chemistry incorporates only one nucleotide at a time using reverse-terminator terminology. This difference presumably explains why the Ion-Torrent platform has significantly higher sequencing error rates compared to the Illumina platform.

Meanwhile, the current v3 sequencing chemistry for MiSeq platform has a larger sequencing data output which is more than 10 Gb of data per run compared to the lower throughput of the Ion-Torrent at ~5.5 Gb (at this stage the up-scale model Ion-Proton instrument is not available commercially yet). The sample preparation protocols for the Illumina platform have a wider variety of choices catering for different types of sample and customer needs. In our project we only utilised both Illumina Nextera and Nextera-XT protocols exploiting the efficiency of enzymatic shearing for metagenomics sample with a cheaper cost, and a better size-selection process (utilizing SPRI beads) with reduced sample handling and preparation time (Lamble, Sarah et al., 2013). However the enzymatic shearing via 'transposase' is not as consistent as we expected as both Nextera (~580bp) and Nextera-XT (~1.2kb) were different in library size even though the samples were the same. A desired library size is very important in Illumina chemistry given it can influence the sequences reads quality and the process of cluster generation (Head et al., 2014). The library size here refers to the size of the target DNA fragments along with the Illumina adapter and index sequences. The clustering process, in which the libraries are denatured, diluted and distributed on the surface of the flowcell prior to amplification and sequencing can determine the quality of the sequence reads. Shorter products tend to amplify more efficiently during bridge-amplification compared to longer products as longer library-inserts generate larger clusters which are more diffuse and this can affect the clustering efficiency (Head et al., 2014). However, we have successfully sequenced the library prepared by the Nextera-XT although it had longer DNA fragments compared to Nextera and both are similar in quality. Also in principle, although paired-end sequencing improves the sequencing coverage it also makes computational analysis more challenging due to an overabundance of many repetitive reads, resulting in more ambiguous assemblies in which it further add complexity towards repetitive homopolymer regions. One suggestion for overcoming this problem is to combine both short

and long reads fragments such as mate-pair (long-read) and paired-end (short-read) libraries or alternatively, a third generation sequencing platform will provide longer reads.

Amongst available technologies, the Illumina platform has become the leading platform of choice for next-generation sequencing and for our project we chose to go ahead with both platforms MiSeq and Ion-Torrent for a comparative investigation. When time is not a factor, then the Illumina platform with multiplexing (batch/barcode/index) on the HiSeq instrument can significant decrease the sequencing cost and provide sequence data not affected by the homopolymer issue.

Low Diversity libraries

One point of consideration concerning the presence of multiple repetitive sequences in our NGS libraries were caused by the low diversity issue. The occurrence of these sequences are known to be problematic for the Illumina sequencing system particularly for the MiSeq platform which is prone to a "matrix/phasing" issue. Phasing here refers to the uneven distribution of DNA nucleotides in a sequenced DNA fragments. Phasing occurs more commonly in the 16S rDNA amplicon due to lower genetic diversity and smaller/narrower DNA fragments thus resulting in an uneven distribution of nucleotides across the flow cell from one cycle to the next which skews the base-calling intensity (Illumina, 2013b). Such low-diversity errors are unique to the Illumina sequencing platform due to its imaging technology and software limitations during the first few cycles of the sequencing process which requires initialization of the sequencing chemistry and imaging. To start the sequencing run, the optics on the Illumina sequencer need to be calibrated on a perfect 'focusing spot' on the tile (focus point), prior to imaging and for low-diversity libraries this image tends to be out of focus causing a miscalculation of cluster density (Paul Barnes, Illumina, private communication, 2012). The template coordinates and focusing spot here refer to the X, Y and Z coordinates of each cluster on a flowcell tile representing the four bases of the DNA nucleotides being imaged four times for cluster intensity and density configuration. Illumina uses two types of lasers - red and green filter wavelength bands - where images of clusters are taken on each cycle with at least a minimum of two nucleotides for each colour channel needed to be read properly before being registered and base-called.

For this reason, when multiplexing low-diversity sample a PhiX library can be spiked into the flowcell to aid the imaging cycles for calibration (Illumina, 2013b).

In the present work, to optimise our prepared Illumina libraries we addressed the matrix/phasing issue by spiking approximately 10% of PhiX library (12.5pM) to counter-balance the genomic DNA base composition across all four nucleotide imaging channels on the MiSeq instrument. This approach is likely to improve conditions for the Illumina in-built computational algorithm (RTA system for de-multiplexing purposes) to call the number of quality reads resulting from alignment and mapping. For analysis we used the Illumina MiSeq RTA software version (1.17.28), which has a number of improvements over earlier software versions: (1) a newer spot-finding algorithm to improve image quality by increasing the lens aperture sensitivity, (2) a new matrix calculation formula allowing the optimal calculation of denser cluster density, (3) a color-coded matrix system where focussing utilizes more imaging cycles (up to 11 cycles in total) to lower the divergence of hazy images and (4) a new phasing correction algorithm where intensity and base-called calculation is now calculated at every 25 cycles instead of each cycle (Illumina, 2013b). According to an Illumina technical note, changes to the newer version of RTA software significantly improved the analysis of low-diversity samples with no changes to any sample preparation workflow.

In our situation spiking of about 10% of a PhiX library into our metagenomics samples did help improve the phasing and the pre-phasing correction score (the score here refers to the uneven nucleotide base coverage and is obtained from the MiSeq reported files). However it is unclear whether spiking of about 1% and 5% returned with the same result and evaluating this requires further investigation.

Read Length

Traditionally, short-read sequences such as those obtained from the Illumina platform are associated with lower genomic-coverage and sequencing bias which can lead to problems of scalability for data assembly and annotation. Recently Illumina have significantly improved and extended the sequencing chemistry and its instrument read length capability. With the recent introduction of version 3 MiSeq sequencing reagent kits with improved chemistry for

higher cluster density and read length which are capable of generating a minimum of 25 million raw sequences reads and 15Gb of output data on a single paired-end 2 x 300 bp run (550 bp insert-size). Paired-end sequencing is where the same set of DNA templates are sequenced twice, one forward and the other reverse. During bioinformatics analysis/assembly of the data, the two sequences can be paired (based on information about the insert size).

This allows greater accuracy in the detection of insertions and deletions (indels), inversions, homopolymers and rearrangements. Having both paired-end reads can significantly improve the read alignment, especially for *de novo* sequencing. With our whole genome sequencing, we produced paired-end data with read lengths of 2 x 151 bp (MiSeq) and 2 x 251 bp (MiSeq). These data were compared with data from the Life-Technologies Ion-Torrent PGM sequencer which was expected to produce 400 bp single-reads.

4.4 Comparison of running costs based on different workflows

The advent of next-generation sequencing and associated cost has been frequently debated as we are now in the era of the $1000 genome (Hayden, 2014). Unfortunately, this situation is only true for researchers undertaking significant amounts of sequencing work i.e. multiple genomes, RNA-seq and exomes sequencing using Illumina high-end instrument such as the HiSeq2000/3000/4000/X-ten platform. However for our project we compared the estimation costs of different library preparation and sequencing platforms. The updated pricing is based on the costing of this project at the time of thesis writing (2014).

There are two major sequencing vendors for our project; Life-Technologies and Illumina. The cost of Ion-Torrent PGM machine is cheaper at $100,000 NZD compared to Illumina MiSeq machine which is about $175,000 NZD. The cost of both instruments are as true as of 2014 from both sequencing vendors.

For library preparation cost, the most expensive reagents per sample preparation is the NEXTFlex DNA PCR-free method at $500 followed closely by Ion-Xpress Plus Fragment Kit at $378, Illumina Nextera at $330 and finally the Illumina Nextera-XT at $150. Next for sequencing running cost the Illumina MiSeq platform with version 2 reagent kit: 2 x 250bp PE run was the priciest at $1,750 per run along with $1572 for a 2 x 150bp PE run.

Meanwhile, the Ion-Torrent PGM 318 chip (400bp) was the cheapest at only about $751 per sequencing run. Overall the most upfront cost for this project was the Nextflex PCR-free protocol on Illumina MiSeq instrument (2 x 250bp) at $2250, followed closely by Illumina Nextera on MiSeq (2 x 150bp) at $1900, Illumina Nextera-XT on MiSeq (2 x 150bp) at $1722 and finally the Ion-Torrent PGM on 318 chip at $1129 (Table 19). As for cost-per-gigabyte of sequencing data, we divided the above sequencing cost to the expected Gb of data from each platform, we concluded the most cost-effective sequencing platform to choose for metagenomics sequencing was using the Illumina MiSeq platform as it only cost about $167 per Gb of data per sample. Thus, at present, the MiSeq instrument remains the most cost-effective platform as it can generate up to 30 million raw-reads with an average yield of 15 Gb (2 x 300 PE run) of paired-end sequencing data.

Summary of actual costing for Tamaki River metagenomics run

Next-generation sequencers	Machine cost ($)	Cost per run ($)	Data output per flow cell	Reads per flow cell	Run Time (Hours)	Cost per GB ($)
Illumina **MiSeq v2**	~175k	~2k	~3-4.5 Gb (2 x 150bp PE) ~6-7.5 Gb (2 x 250bp PE)	~15-20 million reads ~24-30 million reads	~25 ~39	445 267
Illumina **MiSeq v3**	~175k	~2.5k	~12-15 Gb (2 x 300 bp PE)	~44-50 million reads	~48	167
Life-Technologies Ion-Torrent PGM 318 Chip v2	~95k	~750	~1.2-2.0 Gb (400 base reads)	~4-5.5 million reads	~7	375

Table 19 – Summary of specifications of NGS platforms compared in our metagenomics project.

4.5 Computational challenges in our metagenomics analyses.

Currently the computational analyses in the present work needed to address three fundamental questions:

1) What is out there? - How do the taxonomic profiles of microbial populations compare in our metagenomic datasets?
2) How many are they? - Are the microbial organisms in our profiles in similar abundance?
3) What are they doing? - What are the functional attributes of the microbes in our sample?

In our computational analyses, sequencing data was processed according to the following workflow; raw sequencing data were QC checked for sequence quality (FastQC and SolexaQA), sequences were trimmed for any ambiguous data such as sequencing adapter-primer (CutAdapt, DynamicTrim, LengthSort and FastQ-Trimmer), binned, aligned and assembled using BLAST and PAUDA software, before final annotation via MEGAN5 and interpreted using KEGG and SEED classifications.

We compared all the metagenomics data sets for evidence of biases and limitations that might affect taxonomic and attribute conclusions. We encountered problems that included the occurrence of reads from highly repetitive regions which could not be assigned in BLAST searches against the protein database, low quality sequencing reads due to NGS adapter contamination, unassigned singleton reads (more than 5Gb of sequencing data) and lastly high levels of representation of some species in our sequencing data that we might have preferentially amplified them during NGS library construction (PCR biases). Here ‘singleton’ refers to a single read from a single direction that did not assemble or map to a reference sequence. At the moment the singleton reads were unassigned and reserved for future investigation. An example to this approach is a study by Wooley and colleagues explained singleton reads in species-rich samples can be used to infer functional information via short significant BLAST hits (Wooley et al., 2010).

To reduce the poor quality issues with our sequencing reads, the raw sequencing data was first QC filtered (computational tools are described in the Methods and Materials section) prior to being translated into protein sequences for database matching via the PAUDA method before being analysed in MEGAN5. Our analyses indicated that the sequencing data (paired-end reads) was able to generate enough contigs to allow sufficient coverage and annotation for preliminary taxonomic analysis and classification. For instance, most of our assembled contigs were successfully blasted through MEGAN5 using Bowtie2 as the primary mapper. Our results also show that functionally accurate annotation can be achieved from just approximately 1 to 2 million reads (i.e. you do not need a huge amount of dataset for taxa analysis) and this allows characterization of the microbial composition within the collected environmental samples. For ease of communication and to facilitate comparison with earlier studies it is helpful to distinguish short (25 – 250 bp) and long (50 - 400 bp) reads. In general, shorter sequences have been associated with higher accuracy and deeper coverage, whereas longer sequences have been preferred for generating longer contigs, easier assembly and interpretation (Luo et al., 2012; Quail et al., 2012). .

The results obtained in the current work are consistent with earlier reports e.g, (Kircher et al., 2010; Richter et al., 2008) suggesting that analyses of short-read randomly sampled DNA sequences (2x150 and 2 x 250bp paired end data) are sufficient to provide high resolution taxonomic profiles. Longer reads obtained with the Ion-Torrent sequencer were somewhat disappointing, as the data quality were relatively low and suffered from homopolymer errors. However, it is difficult to generalise from this finding on the potential of the Ion-Torrent platform as the sequencing was conducted off site. That is, a detailed investigation was not possible as the Ion-Torrent libraries were made and sequencing undertaken by another NGS sequencing provider (NZGL, Auckland). That said, in respect of differences between the single and paired-end sequencing undertaken, PAUDA and MEGAN analyses of the paired-end Illumina data identified significantly more phylogenetic lineages and taxa than did similar analyses of the Ion-Torrent data. In contrast, biological inferences arising from different Illumina library prep protocols were similar. This means that, at far less cost, similar data could be obtained with an enzymatic (Nextera-XT) protocol without compromising quality.

Storage and handling of metadata

Metagenomics datasets are often complex and can be a challenge for data interpretation due to the huge amount of sequencing data to be processed and often require large and high performance computational resources for genomic analysis. Metagenomic sequencing data should be managed and stored appropriately in a specialised database that includes logical information such as connectivity relationships of microorganisms and the environment. It is intended that the data collected in the present study will be included in a local based metagenomics database network to be established as a bio-monitoring tool for future metagenomics analyses at Massey University. In doing this, for purposes of data storage, we will keep our raw and trimmed reads in fastq format files (removal of any primer/adapter sequences) as read-only files with a Windows 7 permission setting. This is to protect the content within the entire directory against any unwanted changes to the file systems that may cause data corruption. The backup metadata was archived and compressed using the "tape archive" (.Tar) format system. Tar files are a well-known format for maintaining structure integrity. The system file directory identifies structure folders, file permissions, system information and other information. Furthermore, we also compressed and archived the files at the same time thus saving a significant amount of time and storage space. After archiving, the metadata are backed up to several places: one in a working directory for easier retrieval, one in a good-quality external hard-drive and lastly to a cloud storage system such as DropBox for further additional back-up purposes. As we continue with our data analyses, all intermediate analysis files will be continuously backed up to a working tar.gz directory. By maintaining such storage systems we save work, time and cost while using a cloud-based system for more critical data protection.

5 Conclusion

Human population growth means that more attention needs to be given to the sustainability of our drinking water quality. Assessment of freshwater quality will benefit from the future application of cost effective technologies that elucidate the nature of aquatic microbiological ecosystems. Random shotgun metagenomics offers one approach for potential bio-monitoring. Previous gold standard NGS protocols are not cost effective for this purpose. However, here it has been shown that a more time efficient and cost effective protocol that requires far less starting material can be implemented for Illumina NGS metagenomics sequencing. Thus, these results provide a solid foundation for advancing metagenomic studies of aquatic and other ecosystems.

The findings from the present work have already been taken up by researchers in New Zealand through NZGL and as a consequence we can expect to see future studies that advance understanding of native and foreign microbiomes. In particular, this is likely to include studies which investigate the extent to which microbial communities of soils and water interact and diversify. Unlocking the full potential of metagenomics – addressing issues of bias due to database representation, conservative taxonomic assignment and functional diversity of the microbiome remain open challenges for this research (Huson et al., 2016)

6 Future work

This thesis describes the development of NGS methods that can be used to screen and explore freshwater microbiomes. An important goal was to evaluate approaches that could be used in the future to better detect the potential for waterborne disease. The sequencing data from this project are being made publically available and will contribute to a metagenomic database that can be accessed for future comparative analyses.

Currently the most popular metagenomic approaches involves amplicon sequencing (Cottrell, M. T. et al., 2005; Ghai et al., 2011; Savio et al., 2015) and more could be done in future work to compare the findings from amplicon sequencing with results from Nextera-XT genome sequencing. Such studies would be informative in identifying what biological inferences can be reliably drawn from amplicon vs shotgun sequencing NGS approaches. Comparisons could also usefully be made between cDNA sequencing (meta-transcriptomics) (Yu et al., 2012) and Nextera-XT genome sequencing (Trombetta et al., 2014). In this case, for equivalence in gene coverage, deeper levels of Nextera-XT sequencing might be needed for comparisons relating to making inferences of functionality.

Unbiased metagenome sequencing data is important for making reliable biological inferences. One important factor not investigated in the present project are quality controls for library preparation and sequencing consumables. Reagent and laboratory contaminations should to be considered and controlled when using highly sensitive and specific culture-independent techniques. Contamination usually occurs during sample preparation and can have significant impact on the metadata (Motley et al., 2014), which may distort the taxonomic distribution thus affecting the estimation of microbial diversity. Multiple Displacement Amplification (MDA) could be used as one approach for identifying contaminants in reagents (Motley et al., 2014). Previous work has shown that this method is very sensitive to the detection of contaminating DNAs, and thus could be performed on library preparation consumables and sequencing reagents prior to sequencing and taxonomic analyses. The profiles from such a negative control could then be compared with the profile from the environmental study. Further study could also be made on the effectiveness of filters. In the present work, multiple filters were used to maximize DNA yield, however the

selective impact that filter size has on taxonomic inferences was unstudied. This is something that could also be investigated in future work.

Lately advances of NGS platforms such as the introduction of newer sequencing platforms such as Ion-Torrent Proton and Illumina HiSeq3000/4000 and NextSeq500 promise to improve the current NGS data with significant faster, better and cheaper outcomes. Overall the next big improvement towards the current 2nd next-generation sequencer will be based on speed, accuracy and cost-effectiveness. However, newer technology such as the SMRT (Single Molecule Real Time sequencing) or third generation sequencing platforms offer significant advantages for bacterial genomics. Currently, there are two major competitors in this category: Pacific Biosciences (PacBio RS II) and Oxford Nanopore Technologies (MinION, PromethION and GridION). This third generation sequencing technology promises longer read lengths with lower error rates, overcoming hard to sequence reads such as homopolymer regions, cheaper sequencing costs, PCR-bias free and faster turnaround time with highly accurate sequencing reads. With the emergence of these powerful sequencing technologies, future research in freshwater metagenomics could address the following matters:

1) More quantitative approaches for comparing the similarities and differences in microbial profiles (e.g. Huson et al. under review Journal of Biotechnology).
2) Identification of the impact of physical, biogeochemical, and productivity on biotic communities in respect to function and diversity (Jung et al., 2011; Makhalanyane et al., 2016; Peter et al., 2016).
3) Understanding how microbes communicate with each other is an important element for gene regulation and expression. For example research in '*quorum-sensing*' such as by (Ng et al., 2011; Papenfort et al., 2016) signals production and detection specificity in gram-negative bacteria *Vibrio cholerae* provides huge insights of how waterborne pathogenicity and virulence is spread and maintained.
4) Expansion of the current knowledge of how terrestrial ecosystems work in relation to aquatic ecosystems along with understanding the key elements such as the effects micronutrients and macronutrients and human intervention have on biological processes, and how both ecosystems interact with each other in preserving genetic

assimilation and plasticity will greatly enhance our knowledge of complex habitats. (Crook et al., 2015; García-Palacios et al., 2016).

5) Continual improvement of bio-monitoring tools such as recent development of sensing technologies using biological cells as '*bioreporters*' for the detection of biochemical and eco-toxicological activities in environmental contaminants and also
6) Environmental bio-monitoring via targeted amplicon metabarcoding of the highly-conserved Cytochrome C Oxidase subunit I (COI) mitochondrial gene enrichment to revealed taxonomy details without the need of PCR amplification (Dowle et al., 2016; Robbens et al., 2010; Sen et al., 2011).
7) Exploration and review of newer methodology and technology platforms that can increase the efficiency, accuracy and cost-effectiveness of metagenomics applications. For example, the Nanopore Technologies MinION platform promises field sequencing technology with real time result analysis (Karlsson et al., 2015; Wanunu, 2012). Also includes the SMRT (Single Molecule Real Time) sequencing PacBio (Pacific Bioscience) platform. To date, PacBio instrument had been extensively used for microbial genomes for higher coverage and longer sequence reads lenth with higher sequencing accuracy (Koren et al., 2013)
8) Improvements in the assessment of biodiversity by using new novel computational techniques for species identification (Albertsen et al., 2013; Gómez-Zurita et al., 2016; Nakai et al., 2011; Steele et al., 2011; Tanabe et al., 2013).
9) Standardization of computational tools being used on metagenomics bioinformatics pipelines particularly for data-management, storage, integration and analysis (Hanson et al., 2014; Markowitz et al., 2014).
10) Development of rapid inexpensive field deployable isothermal DNA testing which could target microorganisms identified in NGS microbial profiling studies such as undertaken in the present project (Choi et al., 2016).
11) Increasing awareness of the importance of sustaining pollution-free freshwater supplies and their significant impact on the agricultural industry in New Zealand. New regulations may be needed to protect the current freshwater ecosystem. These could be informed by metagenomics studies.

Thus, continued investment into research on aquatic ecosystems is essential to significantly develop bio-monitoring tools and broaden the applications of these powerful technologies.

References

ADAMS, M. D., et al. (2000). The Genome Sequence of Drosophila melanogaster. *Science, 287*(5461), 2185-2195. doi: 10.1126/science.287.5461.2185

AGENCY., U. E. P. (2001). Method 1623: Cryptosporidium and Giardia in water filtration/IMS/FA*., EPA 821-R-01-025.* .

AIRD, D., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol, 12*(2), R18.

ALBERTSEN, M., et al. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol, 31*(6), 533-538. doi: 10.1038/nbt.2579

AMIT ROY, S. R. (2014). Molecular Markers in Phylogenetic Studies-A Review. *Journal of Phylogenetics & Evolutionary Biology, 02*(02). doi: 10.4172/2329-9002.1000131

AMORIM, J. H., et al. (2008). An improved extraction protocol for metagenomic DNA from a soil of the Brazilian Atlantic Rainforest. *Genet Mol Res, 7*(4), 1226-1232.

ANDER, C., et al. (2013). metaBEETL: high-throughput analysis of heterogeneous microbial populations from shotgun DNA sequences. *BMC Bioinformatics, 14 Suppl 5*, S2. doi: 10.1186/1471-2105-14-S5-S2

ANDREW, B. (2006). Estimation of burden of water-borne disease in New Zealand: Preliminary report

ANDREWS, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. doi: citeulike-article-id:11583827

AZAM, F., et al. (2007). Microbial structuring of marine ecosystems. *Nat Rev Microbiol, 5*(10), 782-791. doi: 10.1038/nrmicro1747

BAKER, G. C., et al. (2003). Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods, 55*(3), 541-555.

BAKER, M. G., et al. (2013). Infectious Diseases Attributable to Household Crowding in New Zealand: A systematic review and burden of disease estimate. . *He Kainga Oranga/Housing and Health Research Programme*.

BARZON, L., et al. (2013). Next-generation sequencing technologies in diagnostic virology. *J Clin Virol, 58*(2), 346-350. doi: 10.1016/j.jcv.2013.03.003

BARZON, L., et al. (2011). Applications of next-generation sequencing technologies to diagnostic virology. *Int J Mol Sci, 12*(11), 7861-7884. doi: 10.3390/ijms12117861

BENNETT, S. (2004). Solexa Ltd. *Pharmacogenomics, 5*(4), 433-438. doi: 10.1517/14622416.5.4.433

BERTRAND, H., et al. (2005). High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *Journal of Microbiological Methods, 62*(1), 1-11. doi: DOI: 10.1016/j.mimet.2005.01.003

BIDDLE, J. F., et al. (2011). Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. *Isme Journal*. doi: http://www.nature.com/ismej/journal/vaop/ncurrent/suppinfo/ismej2010199s1.html

BODAKER, I., et al. (2009). Comparative community genomics in the Dead Sea: an increasingly extreme environment. *The ISME Journal, 4*(3), 399-407. doi: citeulike-article-id:6695081

BORGSTRÖM, E., et al. (2011). Large Scale Library Generation for High Throughput Sequencing. *PLoS One, 6*(4), e19119. doi: 10.1371/journal.pone.0019119

BOTTONE, E. J. (1999). Yersinia enterocolitica: overview and epidemiologic correlates. *Microbes and Infection, 1*(4), 323-333. doi: http://dx.doi.org/10.1016/S1286-4579(99)80028-8

BREITBART, M., et al. (2009). Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Ciénegas, Mexico. *Environ Microbiol, 11*(1), 16-34. doi: citeulike-article-id:3862068

BUCHFINK, B., et al. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Meth, 12*(1), 59-60. doi: 10.1038/nmeth.3176

http://www.nature.com/nmeth/journal/v12/n1/abs/nmeth.3176.html#supplementary-information

CANTAREL, B. L., et al. (2011). Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One, 6*(11), e27173. doi: 10.1371/journal.pone.0027173

CARUCCIO, N. (2011). Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. *Methods Mol Biol, 733*, 241-255. doi: 10.1007/978-1-61779-089-8_17

CHAKRAVORTY, S., et al. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods, 69*(2), 330-339. doi: 10.1016/j.mimet.2007.02.005

CHEN, K., et al. (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLoS Computational Biology, 1*(2), e24.

CHOI, J. R., et al. (2016). An integrated paper-based sample-to-answer biosensor for nucleic acid testing at the point of care. *Lab on a Chip, 16*(3), 611-621. doi: 10.1039/C5LC01388G

COTTRELL, M. T., et al. (2005). Bacterial diversity of metagenomic and PCR libraries from the Delaware River. *Environ Microbiol, 7*(12), 1883-1895. doi: 10.1111/j.1462-2920.2005.00762.x

COTTRELL, M. T., et al. (2005). Bacterial diversity of metagenomic and PCR libraries from the Delaware River. *Environ Microbiol, 7*(12), 1883-1895. doi: 10.1111/j.1462-2920.2005.00762.x

COX, M., et al. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics, 11*(1), 485.

CROOK, D. A., et al. (2015). Human effects on ecological connectivity in aquatic ecosystems: Integrating scientific approaches to support management and mitigation. *Science of The Total Environment, 534*, 52-64. doi: http://dx.doi.org/10.1016/j.scitotenv.2015.04.034

DAUGHNEY, C., et al. (2009). National Groundwater Qualiyy Indicators Update: State and Trends 1995-2008. In M. o. ENVIRONMENT (Ed.): Institute of Geological and Nuclear Sciences Limited (GNS Science).

DAVIES, K. (2010). The Solexa Story. *Bio-IT World, October 2010*. doi: http://www.bio-itworld.com/2010/issues/sept-oct/solexa.html

DEL FABBRO, C., et al. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLoS One, 8*(12), e85024. doi: 10.1371/journal.pone.0085024

DENG, X. (2013). Deep Sequencing Applications in Metagenomics. San Francisco.

DJIKENG, A., et al. (2009). Metagenomic Analysis of RNA Viruses in a Fresh Water Lake. *PLoS One, 4*(9), e7264.

DOWLE, E. J., et al. (2016). Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Mol Ecol Resour, 16*(5), 1240-1254. doi: 10.1111/1755-0998.12488

EICHLER, S., et al. (2006). Composition and Dynamics of Bacterial Communities of a Drinking Water Supply System as Assessed by RNA- and DNA-Based 16S rRNA Gene Fingerprinting. *Appl. Environ. Microbiol., 72*(3), 1858-1872. doi: 10.1128/aem.72.3.1858-1872.2006

EWING, B., et al. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res, 8*(3), 186-194.

FALKOWSKI, P. G., et al. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science, 320*(5879), 1034-1039. doi: 10.1126/science.1153213

FANG, J., et al. (2011). Genomics, metagenomics, and microbial oceanography—A sea of opportunities. *SCIENCE CHINA Earth Sciences, 54*(4), 473-480. doi: citeulike-article-id:9158971

FLEISCHMANN, R. D., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science, 269*(5223), 496-512.

FRIAS-LOPEZ, J., et al. (2008). Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences, 105*(10), 3805-3810. doi: 10.1073/pnas.0708897105

GARCÍA-PALACIOS, P., et al. (2016). The importance of litter traits and decomposers for litter decomposition: a comparison of aquatic and terrestrial ecosystems within and across biomes. *Functional Ecology, 30*(5), 819-829. doi: 10.1111/1365-2435.12589

GHAI, R., et al. (2011). Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River. *PLoS One, 6*(8), e23785. doi: 10.1371/journal.pone.0023785

GHAZANFAR, S., et al. (2010). Metagenomics and its application in soil microbial community studies: biotechnological prospects. *Journal of Animal & Plant Sciences, 6*(2), 611-622. doi: citeulike-article-id:8823663

GILBERT, J., et al. (2011). Microbial metagenomics: beyond the genome. *Annual review of marine science, 3*, 347-371. doi: citeulike-article-id:9168262

GÓMEZ-ZURITA, J., et al. (2016). High-throughput biodiversity analysis: Rapid assessment of species richness and ecological interactions of Chrysomelidae (Coleoptera) in the tropics. *ZooKeys*(597), 3-26. doi: 10.3897/zookeys.597.7065

GOTUZZO, E., et al. (1994). Cholera. Lessons from the epidemic in Peru. *Infect Dis Clin North Am, 8*(1), 183-205.

HALL, R. J., et al. (2013). Metagenomic Detection of Viruses in Aerosol Samples from Workers in Animal Slaughterhouses. *PLoS One, 8*(8), e72226. doi: 10.1371/journal.pone.0072226

HANDELSMAN, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiol. Mol. Biol. Rev., 68*(4), 669-685. doi: 10.1128/mmbr.68.4.669-685.2004

HANDELSMAN, J. (2005). Metagenomics or Megagenomics? *Nat Rev Micro, 3*(6), 457-458.

HANSON, N. W., et al. (2014). Metabolic pathways for the whole community. *BMC Genomics, 15*(1), 619. doi: 10.1186/1471-2164-15-619

HASMAN, H., et al. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol, 52*(1), 139-146. doi: 10.1128/jcm.02452-13

HAUSWEDELL, H., et al. (2014). Lambda: the local aligner for massive biological data. *Bioinformatics, 30*(17), i349-i355. doi: 10.1093/bioinformatics/btu439

HAYDEN, E. C. (2014). Is the $1,000 genome for real? *Nature*. doi: 10.1038/nature.2014.14530

HEAD, S. R., et al. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques, 56*(2), 61-64, 66, 68, passim. doi: 10.2144/000114133

HEALTH, M. o. (2015). *Guidelines for Drinking-water Quality Management for New Zealand*.

HERRON, P. R., et al. (1990). New method for extraction of streptomycete spores from soil and application to the study of lysogeny in sterile amended and nonsterile soil. *Appl Environ Microbiol, 56*(5), 1406-1412.

HOLT, R. A., et al. (2008). The new paradigm of flow cell sequencing. *Genome Res, 18*(6), 839-846. doi: 10.1101/gr.073262.107

HUSON, et al. (2007). MEGAN analysis of metagenomic data. *Genome Research, 17*(3), 377-386. doi: 10.1101/gr.5969107

HUSON, et al. (2009). Methods for comparative metagenomics. *BMC Bioinformatics, 10*(Suppl 1), S12.

HUSON, D., et al. (2014). A poor man's BLASTX--high-throughput metagenomic protein database search using PAUDA. *Bioinformatics, 30*(1), 38-39. doi: 10.1093/bioinformatics/btt254

HUSON, D. H., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology, 12*(6), e1004957. doi: 10.1371/journal.pcbi.1004957

ILLUMINA. (2013a). HiSeq 2000 System User Guide. *15011190*(Rev.T), 113. Retrieved from doi:http://supportres.illumina.com/documents/documentation/system_documentation/hiseq2000/hiseq-2000-user-guide-15011190-t.pdf

ILLUMINA. (2013b). Low-Diversity Sequencing on the Illumina MiSeq Platform. from http://www.illumina.com/documents/products/technotes/technote_low_diversity_rta.pdf

ILLUMINA. (2014). Illumina Two-Channel SBS Sequencing Technology. *Illumina Technology Spotlight*. doi: www.illumina.com/systems/nextseq-sequencer/technology.ilmn

JACOBSEN, C. S., et al. (1992). Development and application of a new method to extract bacterial DNA from soil based on separation of bacteria from soil with cation-exchange resin. *Appl Environ Microbiol, 58*(8), 2458-2462.

JANDA, J. M., et al. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol, 45*(9), 2761-2764. doi: 10.1128/JCM.01228-07

JUNG, J. Y., et al. (2011). Metagenomic Analysis of Kimchi, a Traditional Korean Fermented Food. *Appl. Environ. Microbiol., 77*(7), 2264-2274. doi: 10.1128/aem.02157-10

KAKIRDE, K. S., et al. (2010). Size does matter: Application-driven approaches for soil metagenomics. *Soil Biology and Biochemistry, 42*(11), 1911-1923. doi: DOI: 10.1016/j.soilbio.2010.07.021

KANEHISA, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research, 36*(Database issue), D480-D484. doi: 10.1093/nar/gkm882

KARLSSON, E., et al. (2015). Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific Reports, 5*, 11996. doi: 10.1038/srep11996

http://www.nature.com/articles/srep11996#supplementary-information

KELLEY, D. R., et al. (2010). Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics, 11*, 544. doi: 10.1186/1471-2105-11-544

KIM, M., et al. (2013). Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics Inform, 11*(3), 102-113. doi: 10.5808/GI.2013.11.3.102

KIRCHER, M., et al. (2010). High-throughput DNA sequencing – concepts and limitations. *Bioessays, 32*(6), 524-536. doi: 10.1002/bies.200900181

KIRCHMAN, D. L. (2012). *Processes in Microbial Ecology*: OUP Oxford.

KLINDWORTH, A., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research, 41*(1), e1. doi: 10.1093/nar/gks808

KOREN, S., et al. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol, 14*(9), R101-R101. doi: 10.1186/gb-2013-14-9-r101

KOZAREWA, I., et al. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Meth, 6*(4), 291-295. doi: http://www.nature.com/nmeth/journal/v6/n4/suppinfo/nmeth.1311_S1.html

KRSEK, M., et al. (1999). Comparison of different methods for the isolation and purification of total community DNA from soil. *Journal of Microbiological Methods, 39*(1), 1-16.

LAMBLE, S., et al. (2013). Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnology, 13*(1), 1-10. doi: 10.1186/1472-6750-13-104

LAMBLE, S., et al. (2013). Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol, 13*, 104. doi: 10.1186/1472-6750-13-104

LANDER, E. S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi: 10.1038/35057062

LEBRETON, F., et al. (2013). High-Quality Draft Genome Sequence of Vagococcus lutrae Strain LBD1, Isolated from the Largemouth Bass Micropterus salmoides. *Genome Announc, 1*(6). doi: 10.1128/genomeA.01087-13

LECUIT, M., et al. (2014). The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Front Cell Infect Microbiol, 4*, 25. doi: 10.3389/fcimb.2014.00025

LEMARCHAND, K., et al. (2005). Optimization of microbial DNA extraction and purification from raw wastewater samples for downstream pathogen detection by microarrays. *Journal of Microbiological Methods, 63*(2), 115-126. doi: 10.1016/j.mimet.2005.02.021

LIN, S.-K., et al. (2003). Biodiversity of Microbial Life: Foundation of Earth's Biosphere. *Molecules, 8*(2), 223-225.

LINKE, D. (2009). Chapter 34 Detergents: An Overview. In R. B. RICHARD & P. D. MURRAY (Eds.), *Methods in Enzymology* (Vol. Volume 463, pp. 603-617): Academic Press.

LINNARSSON, S. (2010). Recent advances in DNA sequencing methods - general principles of sample preparation. *Experimental Cell Research, 316*(8), 1339-1343. doi: DOI: 10.1016/j.yexcr.2010.02.036

LIU, L., et al. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol, 2012*, 251364. doi: 10.1155/2012/251364

LIU, Z. (2010). *Next Generation Sequencing and Whole Genome Selection in Aquaculture*: Wiley.

LODES, M. (2016). Chimera-Free Library Prep for NGS Platforms | GEN Magazine Articles | GEN. *Vol 32*(No 2). doi: http://www.genengnews.com/gen-articles/chimera-free-library-prep-for-ngs-platforms/3976/

LORENZ, P., et al. (2005). Metagenomics and industrial applications. *Nat Rev Microbiol, 3*(6), 510-516. doi: 10.1038/nrmicro1161

LOUCKS, D. (2005). Fact about Water. *University of Cornell*, 46.

LUO, C., et al. (2012). Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS One, 7*(2), e30087. doi: 10.1371/journal.pone.0030087

MAARIT NIEMI, R., et al. (2001). Extraction and purification of DNA in rhizosphere soil samples for PCR-DGGE analysis of bacterial consortia. *J Microbiol Methods, 45*(3), 155-165.

MAKHALANYANE, T. P., et al. (2016). Microbial diversity and functional capacity in polar soils. *Current Opinion in Biotechnology, 38*, 159-166. doi: http://dx.doi.org/10.1016/j.copbio.2016.01.011

MANICHANH, C., et al. (2008). A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Research, 36*(16), 5180-5188. doi: 10.1093/nar/gkn496

MARDIS, E. R. (2008a). The impact of next-generation sequencing technology on genetics. *Trends in Genetics, 24*(3), 133-141.

MARDIS, E. R. (2008b). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet, 9*, 387 - 402.

MARKOWITZ, V. M., et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research, 42*(Database issue), D560-D567. doi: 10.1093/nar/gkt963

MARQUEZ, G. G. (2002). Water-Borne Diseases: Cholera and Dysentery. *The Complete Idiot's Guide to Dangerous Disease and Epidemics*.

MCCABE, K. M., et al. (1999). Bacterial species identification after DNA amplification with a universal primer pair. *Mol Genet Metab, 66*(3), 205-211. doi: 10.1006/mgme.1998.2795

METZKER, M. (2010). Sequencing technologies - the next generation. *Nat Rev Genet, 11*(1), 31 - 46.

MEYER, F., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9*, 386. doi: 10.1186/1471-2105-9-386

MORE, M. I., et al. (1994). Quantitative cell lysis of indigenous microorganisms and rapid extraction of microbial DNA from sediment. *Appl Environ Microbiol, 60*(5), 1572-1580.

MORGAN, J., et al. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One, 5*(4), e10209. doi: citeulike-article-id:7057608

MOROZOVA, O., et al. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics, 92*(5), 255-264. doi: 10.1016/j.ygeno.2008.07.001

MOTLEY, S. T., et al. (2014). Improved Multiple Displacement Amplification (iMDA) and Ultraclean Reagents. *BMC Genomics, 15*(1), 443. doi: 10.1186/1471-2164-15-443

MURRAY, D. B. a. J. (2008). Direct Isolation of Metagenomic DNA from Environmental Water Samples. *Epicentre Biotechnologies, 15*(1).

MYERS, E. W., et al. (2000). A whole-genome assembly of Drosophila. *Science, 287*(5461), 2196-2204.

NAKAI, R., et al. (2011). Metagenomic Analysis of 0.2-µm-Passable Microorganisms in Deep-Sea Hydrothermal Fluid. *Marine Biotechnology*, 1-9. doi: 10.1007/s10126-010-9351-6

NANNIPIERI, P., et al. (2006). *Nucleic Acids and Proteins in Soil*: Springer.

NEWTON, R. J., et al. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev, 75*(1), 14-49. doi: 10.1128/MMBR.00028-10

NG, W.-L., et al. (2011). Signal production and detection specificity in Vibrio CqsA/CqsS quorum-sensing systems. *Molecular Microbiology, 79*(6), 1407-1417. doi: 10.1111/j.1365-2958.2011.07548.x

ONGLEY, E. D. (1996). *Control of water pollution from agriculture - FAO irrigation and drainage paper 55*. Rome.

PACE, N. R., et al. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc Natl Acad Sci U S A, 109*(4), 1011-1018. doi: 10.1073/pnas.1109716109

PALENIK, B., et al. (2009). Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ Microbiol, 11*(2), 349-359. doi: citeulike-article-id:3939409

PAPENFORT, K., et al. (2016). Quorum sensing signal-response systems in Gram-negative bacteria. *Nat Rev Micro, 14*(9), 576-588. doi: 10.1038/nrmicro.2016.89

PERKINS, T. T., et al. (2013). Choosing a benchtop sequencing machine to characterise Helicobacter pylori genomes. *PLoS One, 8*(6), e67539. doi: 10.1371/journal.pone.0067539

PETER, H., et al. (2016). Shifts in diversity and function of lake bacterial communities upon glacier retreat. *ISME J, 10*(7), 1545-1554. doi: 10.1038/ismej.2015.245

POPE, P. B., et al. (2008). Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *Fems Microbiology Ecology, 64*(1), 9-27. doi: 10.1111/j.1574-6941.2008.00448.x

POPTSOVA, M. S., et al. (2014). Non-random DNA fragmentation in next-generation sequencing. *Scientific Reports, 4*, 4532. doi: 10.1038/srep04532

http://www.nature.com/articles/srep04532#supplementary-information

PRAY, L. A. (2008). Discovery of DNA Structure and Function: Watson and Crick. *Nature, 1*(100). doi: http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397

PRESS, N. A. (2007). *The new science of metagenomics : revealing the secrets of our microbial planet*. Washington, DC: National Academies Press.

QUAIL, M. A., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*, 341. doi: 10.1186/1471-2164-13-341

RACHEL, M., et al. (2014). Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome, 2*(1), 3-3. doi: 10.1186/2049-2618-2-3

RAES, J., et al. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology, 10*(5), 490-498. doi: citeulike-article-id:2914427

RICHTER, D. C., et al. (2008). MetaSim—A Sequencing Simulator for Genomics and Metagenomics. *PLoS One, 3*(10), e3373. doi: 10.1371/journal.pone.0003373

RIESENFELD, C., et al. (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. *Annual Review of Genetics, 38*(1), 525-552. doi: citeulike-article-id:409956

ROBBENS, J., et al. (2010). Escherichia coli as a bioreporter in ecotoxicology. *Applied Microbiology and Biotechnology, 88*(5), 1007-1025. doi: 10.1007/s00253-010-2826-6

ROBE, P., et al. (2003). Extraction of DNA from soil. *European Journal of Soil Biology, 39*(4), 183-190. doi: http://dx.doi.org/10.1016/S1164-5563(03)00033-5

ROCHELLE, P. A., et al. (1997). Comparison of primers and optimization of PCR conditions for detection of Cryptosporidium parvum and Giardia lamblia in water. *Applied and Environmental Microbiology, 63*(1), 106-114.

RONAGHI, M. (2001). Pyrosequencing Sheds Light on DNA Sequencing. *Genome Research, 11*(1), 3-11. doi: 10.1101/gr.150601

RONAGHI, M., et al. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry, 242*(1), 84-89. doi: http://dx.doi.org/10.1006/abio.1996.0432

ROSS, M. G., et al. (2013). Characterizing and measuring bias in sequence data. *Genome Biol, 14*(5), R51. doi: 10.1186/gb-2013-14-5-r51

ROTHBERG, J. M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature, 475*(7356), 348-352. doi: 10.1038/nature10242

RUSCH, D. B., et al. (2007). The "*Sorcerer II*" Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *Plos Biology, 5*(3), e77.

SANGER, F., et al. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology, 162*(4), 729-773. doi: http://dx.doi.org/10.1016/0022-2836(82)90546-0

SANGER, F., et al. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467.

SAVIO, D., et al. (2015). Bacterial diversity along a 2600 km river continuum. *Environ Microbiol, 17*(12), 4994-5007. doi: 10.1111/1462-2920.12886

SCHLEINITZ, K. (2011). Metagenomics – Theory, Methods and Applications. By Diana Marco (Ed.). *Biotechnology Journal, 6*(1), 125-125. doi: 10.1002/biot.201190003

SCHOLZ, M. B., et al. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology, 23*(1), 9-15. doi: http://dx.doi.org/10.1016/j.copbio.2011.11.013

SEN, B., et al. (2011). Transcriptional responses to complex mixtures--A review. *Mutation Research/Reviews in Mutation Research, 636*(1-3), 144-177. doi: 10.1016/j.mrrev.2007.08.002

SHARMA, A. K., et al. (2009). Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater Actinobacteria. *Isme Journal, 3*(6), 726-737. doi: http://www.nature.com/ismej/journal/v3/n6/suppinfo/ismej200913s1.html

SHARON, I., et al. (2009). A Statistical Framework for the Functional Analysis of Metagenomes. In S. BATZOGLOU (Ed.), *Research in Computational Molecular Biology* (Vol. 5541, pp. 496-511): Springer Berlin / Heidelberg.

SHEN, L., et al. (2011). Tween surfactants: Adsorption, self-organization, and protein resistance. *Surface Science, 605*(5-6), 494-499. doi: DOI: 10.1016/j.susc.2010.12.005

SHOKRALLA, S., et al. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology, 21*(8), 1794-1805. doi: 10.1111/j.1365-294X.2012.05538.x

SINGH, R., et al. (2012). Detection and diversity of pathogenic Vibrio from Fiji. *Environ Microbiol Rep, 4*(4), 403-411. doi: 10.1111/j.1758-2229.2012.00344.x

SINSHEIMER, R. L. (1989). The Santa Cruz Workshop—May 1985. *Genomics, 5*(4), 954-956. doi: http://dx.doi.org/10.1016/0888-7543(89)90142-0

SMITH, C., et al. (2012). DNA goes to court. *Nat Biotechnol, 30*(11), 1047-1053. doi: 10.1038/nbt.2408

STEELE, H. L., et al. (2011). Advances in Recovery of Novel Biocatalysts from Metagenomes. *Journal of Molecular Microbiology and Biotechnology, 16*(1-2), 25-37.

STILLER, M., et al. (2009). Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. *Genome Research, 19*(10), 1843-1848. doi: 10.1101/gr.095760.109

STREIT, W. R., et al. (2004). Metagenomics--the key to the uncultured microbes. *Current Opinion in Microbiology, 7*(5), 492-498. doi: citeulike-article-id:409957

SUNNOTEL, O., et al. (2006). Rapid and Sensitive Detection of Single Cryptosporidium Oocysts from Archived Glass Slides. *Journal of Clinical Microbiology, 44*(9), 3285-3291. doi: 10.1128/JCM.00541-06

SWERDLOW, H., et al. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research, 18*(6), 1415-1419. doi: 10.1093/nar/18.6.1415

TANABE, A. S., et al. (2013). Two New Computational Methods for Universal DNA Barcoding: A Benchmark Using Barcode Sequences of Bacteria, Archaea, Animals, Fungi, and Land Plants. *PLoS One, 8*(10), e76910. doi: 10.1371/journal.pone.0076910

TEBBE, C. C., et al. (1993). Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl Environ Microbiol, 59*(8), 2657-2665.

TELESMANICH, N. R., et al. (2011). [Evaluation of toxin producing abilities of non-O1/non-O139 serogroup Vibrio cholerae isolated from humans]. *Zh Mikrobiol Epidemiol Immunobiol*(2), 8-12.

THOMAS, T., et al. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp, 2*(1), 3. doi: 10.1186/2042-5783-2-3

TRINGE, S., et al. (2005). METAGENOMICS: DNA SEQUENCING OF ENVIRONMENTAL SAMPLES. *Nat Rev Genet, 6*(11), 805-814. doi: citeulike-article-id:949609

TROMBETTA, J. J., et al. (2014). Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.], 107*, 4.22.21-24.22.17. doi: 10.1002/0471142727.mb0422s107

TSAI, Y. L., et al. (1991). Rapid method for direct extraction of mRNA from seeded soils. *Appl Environ Microbiol, 57*(3), 765-768.

VAN ZUYLEN, J. (1981). The microscopes of Antoni van Leeuwenhoek. *Journal of Microscopy, 121*(3), 309-328. doi: 10.1111/j.1365-2818.1981.tb01227.x

VENTER, J. C., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi: 10.1126/science.1058040

VENTER, J. C., et al. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74. doi: 10.1126/science.1093857

WANG, J., et al. (2013). Environmental bio-monitoring with high-throughput sequencing. *Brief Bioinform, 14*(5), 575-588. doi: 10.1093/bib/bbt032

WANG, Z., et al. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi: 10.1038/nrg2484

WANUNU, M. (2012). Nanopores: A journey towards DNA sequencing. *Physics of Life Reviews, 9*(2), 125-158. doi: http://dx.doi.org/10.1016/j.plrev.2012.05.010

WHITMAN, W. B., et al. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences, 95*(12), 6578-6583.

WIKIPEDIA, D. (2004). Dannevirke.

WOMMACK, K. E., et al. (2008). Metagenomics: Read Length Matters. *Appl. Environ. Microbiol., 74*(5), 1453-1463. doi: 10.1128/aem.02181-07

WONG, K. H., et al. (2013). Multiplex Illumina Sequencing Using DNA Barcoding *Current Protocols in Molecular Biology*: John Wiley & Sons, Inc.

WOO, P. C. Y., et al. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection, 14*(10), 908-934. doi: 10.1111/j.1469-0691.2008.02070.x

WOOLEY, J. C., et al. (2010). A Primer on Metagenomics. *PLoS Comput Biol, 6*(2), e1000667. doi: 10.1371/journal.pcbi.1000667

WOYKE, T., et al. (2009). Assembling the Marine Metagenome, One Cell at a Time. *PLoS One, 4*(4), e5299. doi: citeulike-article-id:4381405

YU, K., et al. (2012). Metagenomic and Metatranscriptomic Analysis of Microbial Community Structure and Gene Expression of Activated Sludge. *PLoS One, 7*(5), e38183. doi: 10.1371/journal.pone.0038183

ZHANG, H., et al. (2011). Assessment of non-point source pollution using a spatial multicriteria analysis approach. *Ecological Modelling, 222*(2), 313-321. doi: http://dx.doi.org/10.1016/j.ecolmodel.2009.12.011

ZHOU, H.-W., et al. (2011). BIPES, a cost-effective high-throughput method for assessing microbial diversity. *Isme Journal, 5*(4), 741-749. doi: http://www.nature.com/ismej/journal/v5/n4/suppinfo/ismej2010160s1.html

Appendix

(TruSeq DNA Sample Preparation Kit Oligonucleotide Adapter Sequences)

TruSeq Universal Adapter

5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

TruSeq Adapter, Index 1

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 2

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 3

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 4

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 5

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 6

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 7

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 8

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 9

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 10

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 11

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 12

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 13

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 14

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTCCGTATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 15

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACATGTCAGAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 16

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTCCCGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 18

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCCGCACATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 19

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGAAACGATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 20

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGGCCTTATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 21

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTTTCGGAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 22

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTACGTAATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 23

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACGAGTGGATATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 25

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 27

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTCCTTTATCTCGTATGCCGTCTTCTGCTTG

(Nextera DNA Sample Preparation Kit)

Transposon Oligonucleotide Sequences

5′-GCCTCCCTCGCGCCATCAGAGATGTGTATAAGAGACAG

5′-GCCTTGCCAGCCCGCTCAGAGATGTGTATAAGAGACAG

Adapters (showing optional bar code)

5'-AATGATACGGCGACCACCGAGATCTACACGCCTCCCTCGCGCCATCAG

5'-CAAGCAGAAGACGGCATACGAGAT[barcode]CGGTCTGCCTTGCCAGCCCGCTCAG-3'

PCR Primers

5'-AATGATACGGCGACCACCGA

5'-CAAGCAGAAGACGGCATACGA

Oligonucleotide sequences for Genomic DNA

Adapters

5' P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

PCR Primers

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT

Genomic DNA Sequencing Primer

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Oligonucleotide sequences for Paired End DNA

PE Adapters

5' P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

PE PCR Primer 1.0

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

PE PCR Primer 2.0

5' CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

PE Read 1 Sequencing Primer

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

PE Read 2 Sequencing Primer

5' CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Oligonucleotide sequences for the Multiplexing Sample Prep Oligo Only Kit

Multiplexing Adapters

5' P-GATCGGAAGAGCACACGTCT

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Multiplexing PCR Primer 1.0

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Multiplexing PCR Primer 2.0

5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Multiplexing Read 1 Sequencing Primer

5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT

Multiplexing Index Read Sequencing Primer

5' GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Multiplexing Read 2 Sequencing Primer

5' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

PCR Primer, Index 1

5′ CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC

PCR Primer, Index 2

5′ CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTC

PCR Primer, Index 3

5′ CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTC

PCR Primer, Index 4

5′ CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTC

PCR Primer, Index 5

5′ CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTC

PCR Primer, Index 6

5′ CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTC

PCR Primer, Index 7

5′ CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTC

PCR Primer, Index 8

5′ CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTC

PCR Primer, Index 9

5′ CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTC

PCR Primer, Index 10

5′ CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTC

PCR Primer, Index 11

5′ CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTC

PCR Primer, Index 12

5′ CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTC

(Internal Process Controls Oligonucleotide sequences for TruSeq Sample Preparation Kits)

CTE2 - 150bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCTACGTTCCAAATGCAGCGAGCTCGTA
TAACCCTTTAAGAGTTGCTCTTTTTGTTTGGTAAGTTGCAAATCGAAGTTTTAGATTGAGTTCTACGT
CGAGCGGCCGCGAT

CTE2 - 250bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCTTATCTGTCAAAACCGCTAATGTCCG
TTCTAAGACCGTCTGGAGAACACTTGCCCATCAGTGCTTTTGAACCTTTTTTTCACAGGTCCCTTCCG
ATTACACTGAGAAGCTGACCACACCTGCTAGAAGATGGAGGTATGCAGCCCGTTAGTAGGAGTAATAC
TACCCAGCTTATAACCCTCAAACGTAGGGCAGATGGCGGCCGCGAT

CTE2 - 350bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCTAGAGACCATTCGCGATTCCATGAGA

CTCCAAGGGTTCTGCACAACTTATGCACCTCTATTAGATCATTGTGTTCTACGAAGCCTGGACTGCAT
TACATATTCACAACCAACATGAGAAGAGCGGAATAGATGGCCGGATGTTTGGTGGCTTTGATATATTG
TGAGGAGCATTGCGAACCCTAGAGCTGTCCGGTCAAATAACCCCCTCACAATAAGTGTAATGTCATGG
GATAATCAAAAGACTAAGGGAGGGCTTTTATAGAAGGCGTGAGGTCATGCTATCCCCCTCTGAAGACG
CGGCCGCGAT

CTE2 - 450bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCGTATACGTTTCTAATTTGTAGTTAAC
GGTTGGATACCACTTTGAGGCATGTAATATGGTACTGAGCTTCGGCACAGGGCTCAAATTGCATCATT
AAATGTCTCCGATGTGGCTATATGTCATGGATAAAGGCAGCCCCCTATATCTTTTTTTGTGGCAGCAT
GGGTCCATCAAAGCAATTATTCAGGGTCTTAATGACCTCCACAGCTCTAAACGTAATTCATCTGGCTT
TGCCTGTACTTACTTCCTCCATGAAAAAAAGTGTTGATAATGCTCATAATGCTGCCCAGCAATTTCCT
CCCTTCTCAAGACTATTCTGGCTTCCTGGGTACTTAAAAACAGGGCTTAGAGTATGGCTGCTGACAAA
ATTGCACTCTAAACGCTAGCTTAGGTCTTCTGCGGCCGCGAT

CTE2 - 550bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCGTTAGCTATCGTTCGCGAGAAAGTTA
GTAGACACACAGGACCCAGGCGTGCAAGTCAATTTCAGCTGACTACACCGATTCTGGTTAAAAGAGCC
TATGGCCACCCTTATTTTAGAGAAAAAAAACCACACCTCTAATGTGTTGGGCACTAGAAAAAGCTAAC
TACCTAGTCCGTTTCTGGACGACTTCATTGGGAATAACATACCCCCCACTGTGATTAAGACTGGCACT
GTCCTAATGCTTTCTTCAATAGGTTTGGCTCATGTGTGATTCCCTCTGGCAAACTTATAGAGGACAAG
CAGAATAAACCAATTCAAGGTCGTTGTAGCTGAAGGCCTGGCCTGCCTGACAGTTAATTATGAGCATG
TCTTGCCCTTCATGGTGGATATTCACAGCTGAAAGTGGTATTGGCATTTTTTTCTGAGGACACAACGA
GGAAATCTGATAAATACGGCCACCTGAAGTCTAGCTCGGAGTTAACAATTTACCACGTTTAGAGCGGC
CGCGAT

CTE2 - 650bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCGCTCGCACTTAGCCTGTTAAGGGGTT
CGCGCTCGTCTAGTCTGTGCTGTTGCCTGGATAGTAAATTATCATGGTACAAACTTTTAAGAGCCAGT
TAAATGGAGATGGATTTAAAAAGAGTTATTGTAAAGTCTCCCCAGGTGTGTCATTAAATATCCCAACA
GATTGCCCTGGCCTGACCCCCTAAATGCAATTTTGGGATTCCCTTTTAGTTGCTTTCATTAAAATGTA
CCAGCGCAGTAAAAAAAGCACAAAGTATATTGTTTATGTAACTCACTATCTCATTTGCACTGGTTACA
TGGCAGCTTCAGACTGACTAAAACTACACTTTTCCCACCATGGTTCAAAGATCAACAGAACTGGGCCA
ACAAAAGCAATTTTTTCATGTGGTCTAACTACCAACTTATTATGAGTTAAGTTACTTTTAGGTTTAAA
ATCACAGCAGTTTTTCCCTCCACACCTCCCAGAGATACTTTCAGGGTGGCTAAACTTGGCTAAAGGCT
TCCGGACCAACCCTTGTTTCTTTATGGTGCTTGTGTCCTGACAACCGCGTAAGGCATGGAAATTCAGC
TATTTATCCGATCGTTTATATGGGCGTGCGGCCGCGAT

CTE2 - 750bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCTTGGACCGTTAATTCATATATCGAAG

TAGCAGGTTGTTGCCCCGCCTGATGTTGCCACTACTTGCTCATGACAGTTTTTTTAGGCAATGCAAAC
TACTATTTGATATTTTTTTCCAAGTACAGTTGTAGGGTACTCCTTATACTGATTCTTCTGAGCCTGTA
CGGGGAGCATTAGGTACTGATGTAGTAGGAGTTGAGCTTCACAAATTCACCAGGTAAGCCCAAATTTA
TTTTCTGCTTGGACAGGTCCACCTCACATGGGTCTGTCTAATATATTAAAAGAGGGATTTTCTTTGCT
GTATTGCAGCCCAGTATATCTGTTACTTACAGTAGTAGTCCATTATTGCTGGCCTAGGGGCTTTTGCT
CCTACACGAACACCACTCTGTAAAATTTGAGGTCGTCCTTAGAGTCAAACCATTCATGGAGCGCTCTG
TGCATCTACCAACTATCGCTAAGCATTCACTTGGTTGGTTTAAGTGGAGGCAACTCCATTATCTTCTA
GCATACCCTTCCCAGGCTACATGTAGAAAGAGATCTGTTGGGCCCCACTATTTTTTCACCCAGGGAAG
CCTACTTTAGTTATAGCTTGCCAGAGATTTTCTGTGTCATGTAGAAGTCATCCACTTTTAACACCAGG
AGGTGGATGTGGGGCCAGGAAATATGTCAATAACGATACGGGACTTCTAACAGTGACTCGCGGCCGCG
AT

CTE2 - 850bp

ATCCTGCAGATGCATCCAGTACTAGTATGGCCCGGGGGATCCTTAAGTCGTGTCCTTCTCCTACGATC
TTGTGAACGATGGATATTTTCTTTCTAAACTTTAAACAAACAGTGGAGAGATGTTGTTGTGTGTGGAA
CGACGCTTAGCCTACCGAGGAAGATCCAGACTACAATAGAATATGTGGCCAAAACTCTCCGCAACTTC
AGCAGCAAAAAGGATATTATTGACATAACCTCCTCACAAAAAGTACACAAATGGCTAAATAACAGAGC
CCCTCTTTTTACTAGGGAAATGGTGGATGTGGACTTTAGAATTTAAGATAATAAAGCTCTTGATCCCA
ATGTTATTTCCATGTGAGGGACATTAAATTGAGTAACCTTTGCCACATACCCTCTCCCAGAGTCCATT
CTCTAAAACTTGAAGCTCCGCCCCTTTTTACGCACATTAGGCTTCCAATTACGGTCAATGGTCTTGAA
GATTGGGAGCTTTTGAAGAGTAATAAGAACCATCACAAAAAGGAACCCAGAAGCCGGGAGTGTCTACC
AAAAAAATTCAAGGGTTAAAAAAAAGTGACATTTTCTCCTGTTTTTTACACATGATTTTGAATGCTGA
TGGGTCCACGTCCAGCTCTAAAGGTAGGTTCATGGTTCTCCAAAGTTGCTTTCTTGTCAGAATTGAGC
CACATCAGGTAGGTGGGGAAGTAGATCAGTGAGGATGCTTCACATGTGTGGGCACTGGGAACAGAATG
CTTCAATAACACGAGCTGACGAGGGCCCGCTATGAAAAAAAAGATTCTCTGTGCCCCCTGGCGCCTCC
GCACTTAAAGAATTGATGACCGTGCGGCCGCGAT

CTA - 150bp

GGGGGATCCTACGTTCCAAATGCAGCGAGCTCGTATAACCCTTTAAGAGTTGCTCTTTTTGTTTGGTA
AGTTGCAAATCGAAGTTTTAGATTGAGTTCTACGTCGAGCGGCCGCGATATCCTGCAGATGCATCCAG
TACTAGTATGGCCC

CTA - 250bp

GGGGGATCCTTATCTGTCAAAACCGCTAATGTCCGTTCTAAGACCGTCTGGAGAACACTTGCCCATCA
GTGCTTTTGAACCTTTTTTTCACAGGTCCCTTCCGATTACACTGAGAAGCTGACCACACCTGCTAGAA
GATGGAGGTATGCAGCCCGTTAGTAGGAGTAATACTACCCAGCTTATAACCCTCAAACGTAGGGCAGA
TGGCGGCCGCGATATCCTGCAGATGCATCCAGTACTAGTATGGCCC

CTA - 350bp

GGGGGATCCTAGAGACCATTCGCGATTCCATGAGACTCCAAGGGTTCTGCACAACTTATGCACCTCTA

TTAGATCATTGTGTTCTACGAAGCCTGGACTGCATTACATATTCACAACCAACATGAGAAGAGCGGAA
TAGATGGCCGGATGTTTGGTGGCTTTGATATATTGTGAGGAGCATTGCGAACCCTAGAGCTGTCCGGT
CAAATAACCCCCTCACAATAAGTGTAATGTCATGGGATAATCAAAAGACTAAGGGAGGGCTTTTATAG
AAGGCGTGAGGTCATGCTATCCCCCTCTGAAGACGCGGCCGCGATATCCTGCAGATGCATCCAGTACT
AGTATGGCCC

CTA - 450bp

GGGGGATCCGTATACGTTTCTAATTTGTAGTTAACGGTTGGATACCACTTTGAGGCATGTAATATGGT
ACTGAGCTTCGGCACAGGGCTCAAATTGCATCATTAAATGTCTCCGATGTGGCTATATGTCATGGATA
AAGGCAGCCCCCTATATCTTTTTTTGTGGCAGCATGGGTCCATCAAAGCAATTATTCAGGGTCTTAAT
GACCTCCACAGCTCTAAACGTAATTCATCTGGCTTTGCCTGTACTTACTTCCTCCATGAAAAAAAGTG
TTGATAATGCTCATAATGCTGCCCAGCAATTTCCTCCCTTCTCAAGACTATTCTGGCTTCCTGGGTAC
TTAAAAACAGGGCTTAGAGTATGGCTGCTGACAAAATTGCACTCTAAACGCTAGCTTAGGTCTTCTGC
GGCCGCGATATCCTGCAGATGCATCCAGTACTAGTATGGCCC

CTA - 550bp

GGGGGATCCGTTAGCTATCGTTCGCGAGAAAGTTAGTAGACACACAGGACCCAGGCGTGCAAGTCAAT
TTCAGCTGACTACACCGATTCTGGTTAAAAGAGCCTATGGCCACCCTTATTTTAGAGAAAAAAAACCA
CACCTCTAATGTGTTGGGCACTAGAAAAAGCTAACTACCTAGTCCGTTTCTGGACGACTTCATTGGGA
ATAACATACCCCCCACTGTGATTAAGACTGGCACTGTCCTAATGCTTTCTTCAATAGGTTTGGCTCAT
GTGTGATTCCCTCTGGCAAACTTATAGAGGACAAGCAGAATAAACCAATTCAAGGTCGTTGTAGCTGA
AGGCCTGGCCTGCCTGACAGTTAATTATGAGCATGTCTTGCCCTTCATGGTGGATATTCACAGCTGAA
AGTGGTATTGGCATTTTTTTCTGAGGACACAACGAGGAAATCTGATAAATACGGCCACCTGAAGTCTA
GCTCGGAGTTAACAATTTACCACGTTTAGAGCGGCCGCGATATCCTGCAGATGCATCCAGTACTAGTA
TGGCCC

CTA - 650bp

GGGGGATCCGCTCGCACTTAGCCTGTTAAGGGGTTCGCGCTCGTCTAGTCTGTGCTGTTGCCTGGATA
GTAAATTATCATGGTACAAACTTTTAAGAGCCAGTTAAATGGAGATGGATTTAAAAAGAGTTATTGTA
AAGTCTCCCCAGGTGTGTCATTAAATATCCCAACAGATTGCCCTGGCCTGACCCCCTAAATGCAATTT
TGGGATTCCCTTTTAGTTGCTTTCATTAAAATGTACCAGCGCAGTAAAAAAAGCACAAAGTATATTGT
TTATGTAACTCACTATCTCATTTGCACTGGTTACATGGCAGCTTCAGACTGACTAAAACTACACTTTT
CCCACCATGGTTCAAAGATCAACAGAACTGGGCCAACAAAAGCAATTTTTTCATGTGGTCTAACTACC
AACTTATTATGAGTTAAGTTACTTTTAGGTTTAAAATCACAGCAGTTTTTCCCTCCACACCTCCCAGA
GATACTTTCAGGGTGGCTAAACTTGGCTAAAGGCTTCCGGACCAACCCTTGTTTCTTTATGGTGCTTG
TGTCCTGACAACCGCGTAAGGCATGGAAATTCAGCTATTTATCCGATCGTTTATATGGGCGTGCGGCC
GCGATATCCTGCAGATGCATCCAGTACTAGTATGGCCC

CTA - 750bp

GGGGGATCCTTGGACCGTTAATTCATATATCGAAGTAGCAGGTTGTTGCCCCGCCTGATGTTGCCACT
ACTTGCTCATGACAGTTTTTTTAGGCAATGCAAACTACTATTTGATATTTTTTTCCAAGTACAGTTGT
AGGGTACTCCTTATACTGATTCTTCTGAGCCTGTACGGGGAGCATTAGGTACTGATGTAGTAGGAGTT
GAGCTTCACAAATTCACCAGGTAAGCCCAAATTTATTTTCTGCTTGGACAGGTCCACCTCACATGGGT
CTGTCTAATATATTAAAAGAGGGATTTTCTTTGCTGTATTGCAGCCCAGTATATCTGTTACTTACAGT
AGTAGTCCATTATTGCTGGCCTAGGGGCTTTTGCTCCTACACGAACACCACTCTGTAAAATTTGAGGT
CGTCCTTAGAGTCAAACCATTCATGGAGCGCTCTGTGCATCTACCAACTATCGCTAAGCATTCACTTG
GTTGGTTTAAGTGGAGGCAACTCCATTATCTTCTAGCATACCCTTCCCAGGCTACATGTAGAAAGAGA
TCTGTTGGGCCCCACTATTTTTTCACCCAGGGAAGCCTACTTTAGTTATAGCTTGCCAGAGATTTTCT
GTGTCATGTAGAAGTCATCCACTTTTAACACCAGGAGGTGGATGTGGGGCCAGGAAATATGTCAATAA
CGATACGGGACTTCTAACAGTGACTCGCGGCCGCGATATCCTGCAGATGCATCCAGTACTAGTATGGC
CC

CTA - 850bp

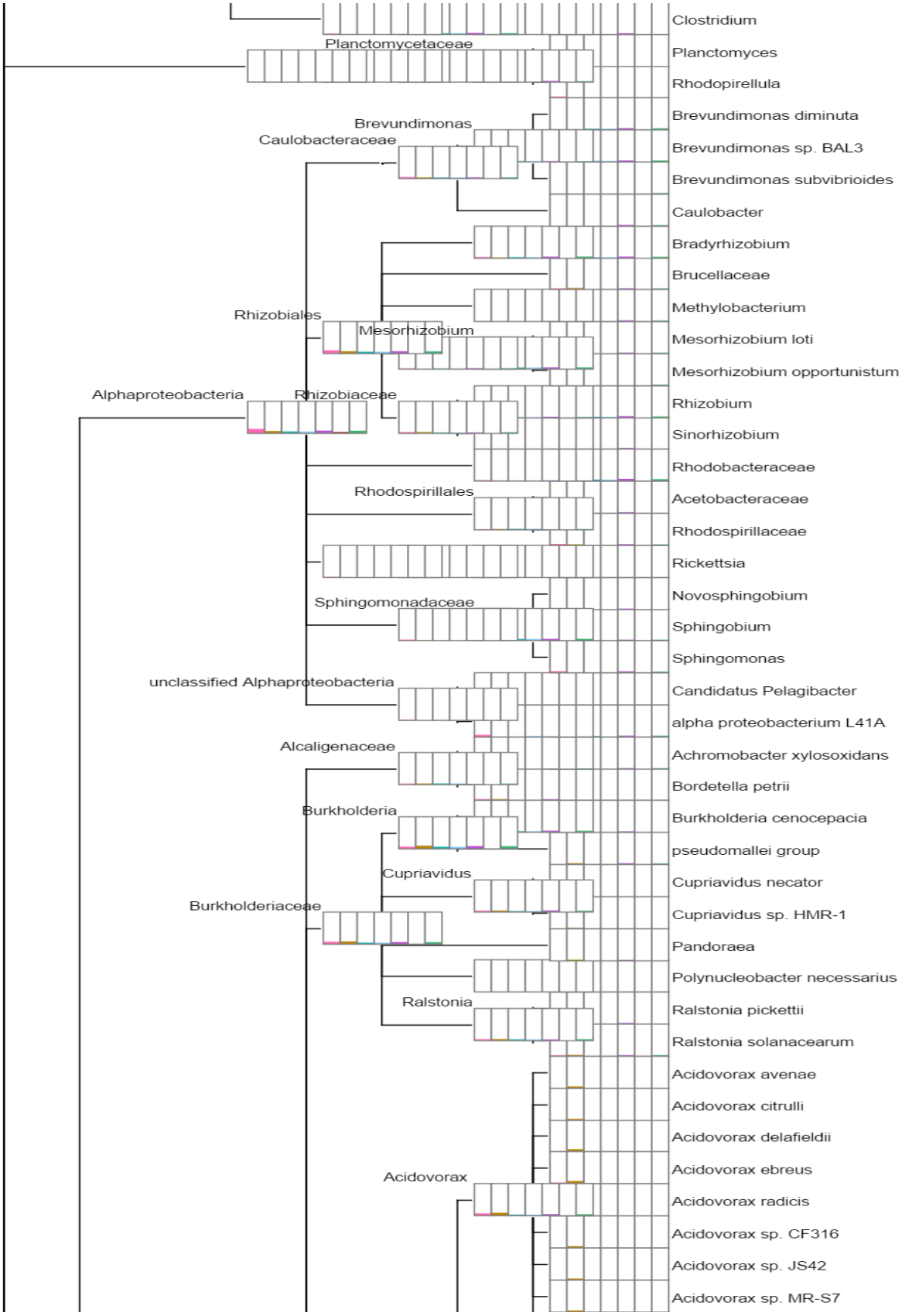
GGGGGATCCTTAAGTCGTGTCCTTCTCCTACGATCTTGTGAACGATGGATATTTTCTTTCTAAACTTT
AAACAAACAGTGGAGAGATGTTGTTGTGTGTGGAACGACGCTTAGCCTACCGAGGAAGATCCAGACTA
CAATAGAATATGTGGCCAAAACTCTCCGCAACTTCAGCAGCAAAAAGGATATTATTGACATAACCTCC
TCACAAAAAGTACACAAATGGCTAAATAACAGAGCCCCTCTTTTTACTAGGGAAATGGTGGATGTGGA
CTTTAGAATTTAAGATAATAAAGCTCTTGATCCCAATGTTATTTCCATGTGAGGGACATTAAATTGAG
TAACCTTTGCCACATACCCTCTCCCAGAGTCCATTCTCTAAAACTTGAAGCTCCGCCCCTTTTTACGC
ACATTAGGCTTCCAATTACGGTCAATGGTCTTGAAGATTGGGAGCTTTTGAAGAGTAATAAGAACCAT
CACAAAAAGGAACCCAGAAGCCGGGAGTGTCTACCAAAAAAATTCAAGGGTTAAAAAAAAGTGACATT
TTCTCCTGTTTTTTACACATGATTTTGAATGCTGATGGGTCCACGTCCAGCTCTAAAGGTAGGTTCAT
GGTTCTCCAAAGTTGCTTTCTTGTCAGAATTGAGCCACATCAGGTAGGTGGGGAAGTAGATCAGTGAG
GATGCTTCACATGTGTGGGCACTGGGAACAGAATGCTTCAATAACACGAGCTGACGAGGGCCCGCTAT
GAAAAAAAAGATTCTCTGTGCCCCCTGGCGCCTCCGCACTTAAAGAATTGATGACCGTGCGGCCGCGA
TATCCTGCAGATGCATCCAGTACTAGTATGGCCC

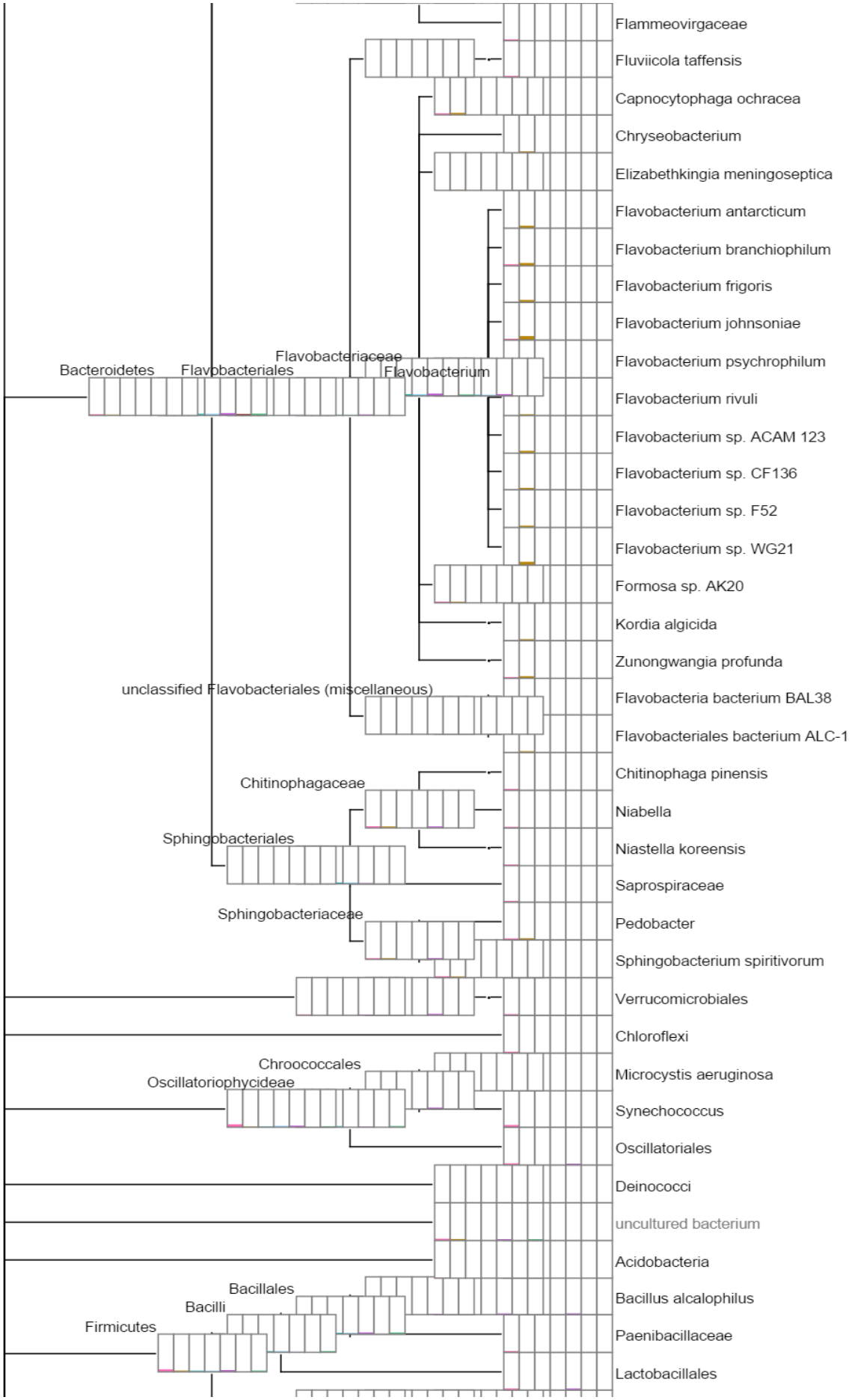
(NextFlex PCR-free kit DNA Adapter Oligonucleotide Sequences, Illumina Compatible)

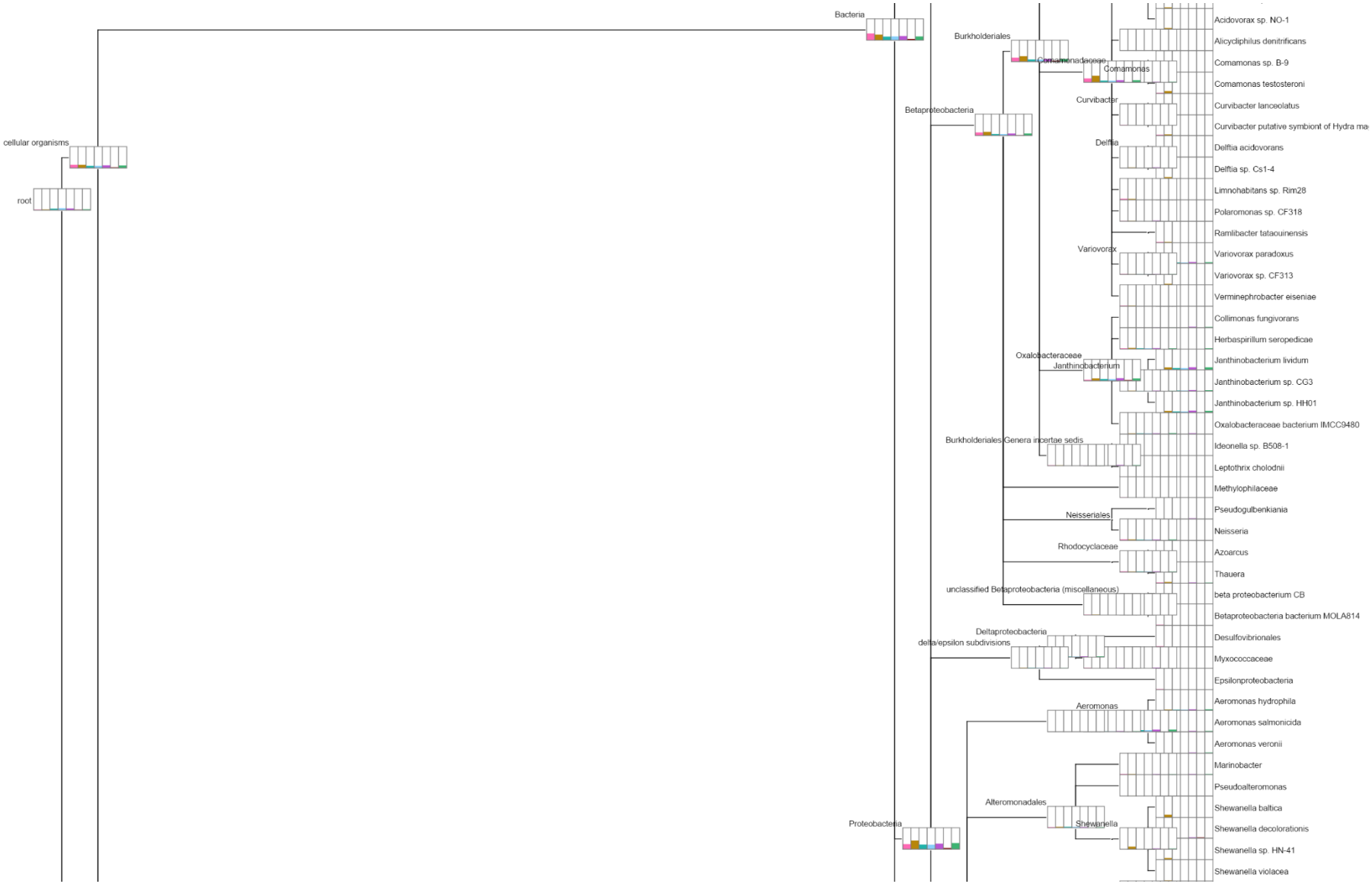
5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

5'GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG

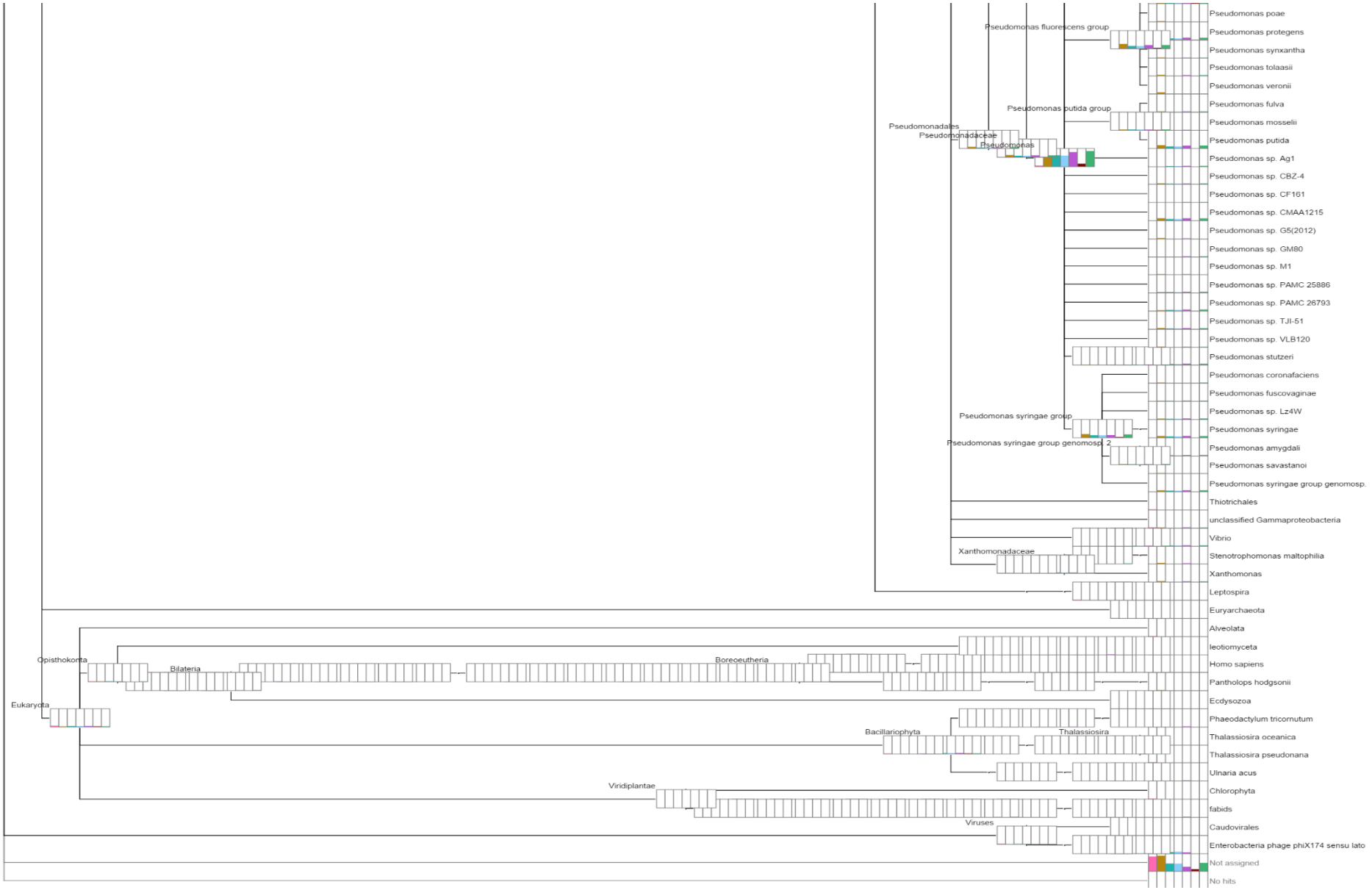
Complete taxonomy profiles showing bacteria species found in our metagenomics datasets generated via Megan5 software




Flammeovirgaceae
Fluviicola taffensis
Capnocytophaga ochracea
Chryseobacterium
Elizabethkingia meningoseptica
Flavobacterium antarcticum
Flavobacterium branchiophilum
Flavobacterium frigoris
Flavobacterium johnsoniae
Flavobacterium psychrophilum
Flavobacterium rivuli
Flavobacterium sp. ACAM 123
Flavobacterium sp. CF136
Flavobacterium sp. F52
Flavobacterium sp. WG21
Formosa sp. AK20
Kordia algicida
Zunongwangia profunda
Flavobacteria bacterium BAL38
Flavobacteriales bacterium ALC-1
Chitinophaga pinensis
Niabella
Niastella koreensis
Saprospiraceae
Pedobacter
Sphingobacterium spiritivorum
Verrucomicrobiales
Chloroflexi
Microcystis aeruginosa
Synechococcus
Oscillatoriales
Deinococci
uncultured bacterium
Acidobacteria
Bacillus alcalophilus
Paenibacillaceae
Lactobacillales
Bacteroidetes
Flavobacteriales
Flavobacteriaceae
Flavobacterium
unclassified Flavobacteriales (miscellaneous)
Chitinophagaceae
Sphingobacteriales
Sphingobacteriaceae
Chroococcales
Oscillatoriophycideae
Bacillales
Bacilli
Firmicutes


cellular organisms
root
Bacteria
Proteobacteria
Betaproteobacteria
Burkholderiales
Comamonadaceae
Comamonas
Curvibacter
Delftia
Variovorax
Oxalobacteraceae
Janthinobacterium
Burkholderiales Genera incertae sedis
Neisseriales
Rhodocyclaceae
unclassified Betaproteobacteria (miscellaneous)
Deltaproteobacteria
delta/epsilon subdivisions
Aeromonas
Alteromonadales
Shewanella
Acidovorax sp. NO-1
Alicycliphilus denitrificans
Comamonas sp. B-9
Comamonas testosteroni
Curvibacter lanceolatus
Curvibacter putative symbiont of Hydra ma
Delftia acidovorans
Delftia sp. Cs1-4
Limnohabitans sp. Rim28
Polaromonas sp. CF318
Ramlibacter tataouinensis
Variovorax paradoxus
Variovorax sp. CF313
Verminephrobacter eiseniae
Collimonas fungivorans
Herbaspirillum seropedicae
Janthinobacterium lividum
Janthinobacterium sp. CG3
Janthinobacterium sp. HH01
Oxalobacteraceae bacterium IMCC9480
Ideonella sp. B508-1
Leptothrix cholodnii
Methylophilaceae
Pseudogulbenkiania
Neisseria
Azoarcus
Thauera
beta proteobacterium CB
Betaproteobacteria bacterium MOLA814
Desulfovibrionales
Myxococcaceae
Epsilonproteobacteria
Aeromonas hydrophila
Aeromonas salmonicida
Aeromonas veronii
Marinobacter
Pseudoalteromonas
Shewanella baltica
Shewanella decolorationis
Shewanella sp. HN-41
Shewanella violacea


Chromatiaceae
Citrobacter
Dickeya
Enterobacter
Enterobacter cloacae complex
Enterobacter sp. 638
Erwinia
Escherichia coli
Klebsiella pneumoniae
Pantoea
Pectobacterium carotovorum
Photorhabdus
Proteus
Providencia
Enterobacteriaceae
Rahnella aquatilis
Salmonella enterica
Serratia fonticola
Serratia marcescens
Serratia
Serratia plymuthica
Serratia proteamaculans
Serratia sp. DD3
unclassified Enterobacteriaceae (miscell
Yersinia aldovae
Yersinia enterocolitica
Yersinia
Yersinia intermedia
Yersinia kristensenii
Yersinia pestis
Yersinia pseudotuberculosis
Legionella pneumophila
Methylococcaceae
Halomonas
Pasteurellacaae
Acinetobacter baumannii
Gammaproteobacteria
Azotobacter vinelandii
Pseudomonas aeruginosa
Pseudomonas aeruginosa group
Pseudomonas alcaligenes
Pseudomonas mendocina
Pseudomonas pseudoalcaligenes
Pseudomonas brassicacearum
Pseudomonas chlororaphis group
Pseudomonas chlororaphis
Pseudomonas fragi
Pseudomonas entomophila
Pseudomonas extremaustralis
Pseudomonas fluorescens


Pseudomonas poae
Pseudomonas protegens
Pseudomonas fluorescens group
Pseudomonas synxantha
Pseudomonas tolaasii
Pseudomonas veronii
Pseudomonas fulva
Pseudomonas putida group
Pseudomonas mosselii
Pseudomonadales
Pseudomonadaceae
Pseudomonas
Pseudomonas putida
Pseudomonas sp. Ag1
Pseudomonas sp. CBZ-4
Pseudomonas sp. CF161
Pseudomonas sp. CMAA1215
Pseudomonas sp. G5(2012)
Pseudomonas sp. GM80
Pseudomonas sp. M1
Pseudomonas sp. PAMC 25886
Pseudomonas sp. PAMC 26793
Pseudomonas sp. TJI-51
Pseudomonas sp. VLB120
Pseudomonas stutzeri
Pseudomonas coronafaciens
Pseudomonas fuscovaginae
Pseudomonas sp. Lz4W
Pseudomonas syringae group
Pseudomonas syringae
Pseudomonas syringae group genomosp. 2
Pseudomonas amygdali
Pseudomonas savastanoi
Pseudomonas syringae group genomosp.
Thiotrichales
unclassified Gammaproteobacteria
Vibrio
Xanthomonadaceae
Stenotrophomonas maltophilia
Xanthomonas
Leptospira
Euryarchaeota
Alveolata
Opisthokonta
Bilateria
Boreoeutheria
leotiomyceta
Homo sapiens
Pantholops hodgsonii
Eukaryota
Ecdysozoa
Phaeodactylum tricornutum
Bacillariophyta
Thalassiosira
Thalassiosira oceanica
Thalassiosira pseudonana
Ulnaria acus
Viridiplantae
Chlorophyta
fabids
Viruses
Caudovirales
Enterobacteria phage phiX174 sensu lato
Not assigned
No hits

Complete SEED analysis of all metagenomics datasets


SEED
Dormancy and Sporulation
Amino Acids and Derivatives
Iron acquisition and metabolism
Phages, Prophages, Transposable elemen
Cell Division and Cell Cycle
Membrane Transport
Nitrogen Metabolism
Sulfur Metabolism
DNA Metabolism
Cofactors, Vitamins, Prosthetic Groups, Pig
Fatty Acids, Lipids, and Isoprenoids
Virulence, Disease and Defense
Potassium metabolism
Motility and Chemotaxis
Virulence
Not assigned
No hits

Complete KEGG analysis of all metagenomics datasets